



**Scuola Internazionale Superiore di Studi Avanzati  
Trieste**

## More Equal than Others

The neural basis of unfairness and inequality perception in  
the Ultimatum Game

Candidate:

Claudia Civali

Supervisor:

Raffaella Rumiati

Thesis submitted for the degree of Doctor of Philosophy in  
Cognitive Neuroscience

Trieste, 2011

**SISSA - Via Bonomea 265 - 34136 TRIESTE - ITALY**

# **Jury**

## **Stefano Cappa**

Dipartimento di Neuroscienze,  
Universita' Vita-Salute San Raffaele, Milano, Italy.

## **Mathew Diamond**

Cognitive Neuroscience Sector,  
International School for Advanced Studies (SISSA/ISAS), Trieste, Italy.

## **Raffaella Rumiati**

Cognitive Neuroscience Sector,  
International School for Advanced Studies (SISSA/ISAS), Trieste, Italy.

## **Alan Sanfey**

Donders Center for Cognitive Neuroscience,  
Radboud University, Nijmegen, the Netherlands.

## **Giorgia Silani**

Cognitive Neuroscience Sector,  
International School for Advanced Studies (SISSA/ISAS), Trieste, Italy.

## **Giosue' Baggio (substitute)**

Cognitive Neuroscience Sector,  
International School for Advanced Studies (SISSA/ISAS), Trieste, Italy.

# Acknowledgments

First and foremost, I would like to thank Raffaella Rumiati, my supervisor, and Corrado Corradi Dell'Aqua, aka CCD, whom I have collaborated with, and informally co-supervised most of the present work. I thank Raffaella for having always believed in me and in my ideas, giving me both the chance and the means I needed to develop them. I thank CCD for being such a knowledgeable and patient teacher, as well as a good friend, and for always being there for me despite all our little arguments! Thank you, this work is ours. I would also like to thank Aldo Rustichini, Cristiano Crescentini and Mathias Gamer for having accepted to give their precious contributions to my work.

I could not have made it without all my wonderful lab/dept. mates and friends, who made my time here unforgettable, broadening the horizons of my knowledge and making me feel part of something great, which SISSA is. Special thanks to Valentina Daelli, Eleonora Russo, Federico Stella, Olga Puccioni, and Paola Mengotti. Last but not least, thanks to Andrea Sciarrone and Alessio Isaja, the backbones of the Sector: none of us would be here without them!

I am very grateful to SISSA for having given me the opportunity to move to Trieste and to meet those who turned out to be some of the most important people in my life, like Liuba Papeo, Antonella Masetto, Adriano Rosso, Michelangelo Mongiello, and Sara Bellinato. Thank you guys for making my life better every day.

I thank my –extended- family (Ali, Vero, Fabri, Clelia, Alberto, Maddalena, Anedi, Nonna, Teo, Micky, Massi, Pelo, Ska, and Eli) and, especially, my mum and my dad, for having been supporting me, in any possible way, since ever, and never getting bored of doing that. You are incredible.

Finally, I thank Carlo, because he has chosen to take the plunge and share his future with mine, teaching me what trust is. Love.

# Contents

<b>Abstract.....</b>	<b>6</b>
<b>Chapter 1: Introduction .....</b>	<b>9</b>
1.1 Different approaches to study decisions .....	10
1.1.1 <i>Game Theory</i> .....	10
1.1.2 <i>Behavioral Economics: social preferences and bounded rationality</i> .....	13
1.1.3 <i>The contribution of Cognitive and Brain Science and the rise of Neuroeconomics</i> ....	14
1.2 The Ultimatum Game .....	17
1.2.1 <i>Theories of social preferences</i> .....	18
1.2.2 <i>Psychophysiological studies: skin conductance response (SCR)</i> .....	19
1.2.3 <i>Brain imaging studies: functional magnetic resonance (fMRI)</i> .....	20
1.2.4 <i>Neuropsychological studies: prefrontal patients</i> .....	22
1.2.5 <i>Brain stimulation studies: transcranial magnetic stimulation (TMS), transcranial direct current stimulation (tDCS), tryptophan depletion</i> .....	23
1.2.6 <i>What is there in a rejection? Open questions and hypothesis</i> .....	24
 <b>Chapter 2: Are irrational reactions to unfairness truly emotionally-driven? Dissociated behavioral and emotional responses.....</b>	 <b>25</b>
2.1 Introduction .....	25
2.2 Methods .....	27
2.2.1 <i>Participants</i> .....	27
2.2.2 <i>Task</i> .....	27
2.2.2 <i>Apparatus and Procedure</i> .....	30
2.2.3 <i>Skin Conductance Recordings</i> .....	30
2.2.4 <i>Emotional Ratings</i> .....	31
2.3 Results .....	31
2.3.1 <i>Rejection Rate</i> .....	32
2.3.2 <i>Emotional Ratings</i> .....	33
2.3.4 <i>Skin Conductance Response amplitude</i> .....	34
2.4 Discussion .....	36

<b>Chapter 3: Disentangling self- and fairness- related neural mechanisms: an fMRI study ...</b>	<b>41</b>
3.1 Introduction .....	41
3.2 Methods .....	43
3.2.1 <i>Participants</i> .....	43
3.2.2 <i>Task and stimuli</i> .....	43
3.2.2 <i>Experimental Set-Up</i> .....	44
3.3.4 <i>Behavioral and imaging data processing</i> .....	45
3.3 Results .....	48
3.3.1 <i>Behavioral results</i> .....	48
3.3.2 <i>Neural Activations</i> .....	49
3.4 Discussion .....	57
3.4.1 <i>Self-Specific Neural Networks</i> .....	57
3.4.2 <i>Fairness-Related Neural Networks</i> .....	61
 <b>Chapter 4: Driving principles in decision-making: the role of abstract social rules .....</b>	 <b>66</b>
4.1 Introduction .....	66
4.1.1 <i>Experimental method</i> .....	70
4.2 Experiment 1: third party UG with an external proposer .....	71
4.2.1 <i>Method</i> .....	72
4.2.2 <i>Results and discussion</i> .....	73
4.3 Experiment 2: the UG with allocators' manipulation .....	75
4.3.1 <i>Method</i> .....	77
4.3.2 <i>Results and discussion</i> .....	78
4.4 Experiment 3: third-party UG with coin's allocations .....	79
4.4.1 <i>Method</i> .....	80
4.4.2 <i>Results and discussion</i> .....	82
4.5 General discussion .....	84
4.5.1 <i>Negative reciprocity</i> .....	85
4.5.2 <i>Inequality aversion</i> .....	87
4.5.3 <i>Self-involvement</i> .....	89

<b>Chapter 5: The neural basis of inequality as an abstract social rule</b>	<b>92</b>
5.1 Introduction	92
5.2 Methods	94
5.2.1 <i>Participants</i>	94
5.2.2 <i>Task and stimuli</i>	94
5.2.3 <i>fMRI data acquisition and experimental set-up</i>	97
5.2.4 <i>Behavioral and imaging data processing</i>	98
5.3 Results	99
5.3.1 <i>Behavioral results</i>	100
5.3.2 <i>Neural Activations</i>	100
5.4 Discussion	107
5.4.1 <i>Self-Specific Neural Networks</i>	108
5.4.2 <i>Fairness-Related Neural Networks</i>	110
 <b>Chapter 6: General discussion</b>	 <b>112</b>
6.1 Emotional involvement and self-concerns	114
6.2 Negative reciprocity, inequality aversion and self-involvement as salient contextual cue	115
6.3 Inequality aversion as a default mode: moral norm or social heuristic?	118
6.4 The role of the anterior insula	120
6.5 Concluding remarks	121
 <b>References</b>	 <b>123</b>

# Abstract

As the research on the Ultimatum Game (UG) has clearly demonstrated, the model of *homo economicus* fails to predict human behavior in a number of situations.. Many interpretations have been put forward in order to explain why players do not simply aim at maximizing their monetary payoff. For instance, models of social preferences (see e.g., Camerer, 2003) try to provide a formal explanation for the apparently irrational behavior of people facing a certain kind of interactive situation. In **chapter 1**, a more detailed description of these accounts will be given, focusing especially on the specific case of the UG in relation to negative reciprocity (Rabin, 1993) and inequity aversion (Fehr & Schmidt, 1999). From a psychological viewpoint, negative emotions, such as anger and frustration, elicited by the unfair treatment, are accounted to cause rejections (Pillutla and Murnighan, 1996).

Neuroscientific findings support the idea that emotional reactions to unfairness cause the deviation from the rational expectation (van't Wout et al., 2006; Sanfey et al., 2003); however, in its traditional formulation, the UG is a self-centered task: the offers address only the responder, and so does the potential unfairness. Therefore, it is impossible to say whether the frustration is elicited by the pure perception of unfairness or by the fact that the unfairness is affecting directly the responder, damaging his/her own payoff. In **chapters 2-3**, I will describe the modified version of the UG I have developed: asking participants to play as responders in both a classical UG (UG\_MS –myself- condition) and a third party UG (UG\_TP), where they had to decide to accepting or rejecting offers on behalf of the next responder, allows disentangling between the two possible options, i.e. frustration elicited by the perception of pure unfairness, or frustration elicited by the self-directed nature of the unfairness. Skin conductance response (SCR) (chapter

2) and BOLD signal (chapter 3) are used as indexes of emotional arousal and neural activation, respectively. In both studies, participants' behavior shows no difference in UG\_MS and in UG\_TP (specifically, rejection rate is higher for unfair offers, and decreases as the offers become fairer both in UG\_MS and in UG\_TP); however, behavior dissociates both from the psychophysiological and the neural evidence. In particular, participants are more aroused (higher SCR and subjective emotional ratings) when rejecting compared to accepting offers in UG\_MS, but not in UG\_TP, where, instead, there is no effect of response; the medial prefrontal cortex (MPFC), an area which has been associated to emotions and self-perception, shows an increase of activation when rejecting, as opposed to accepting, offers in UG\_MS, but not in the UG\_TP, whereas a higher activation in anterior insula (AI) is associated with rejections for both MS and TP. The results from these two studies suggest that, albeit emotions clearly enter the decision-making process, such as in MS, they should not be held as being the only mechanism that triggers rejections.

In the traditional version of the UG, it is unclear whether the aversion towards low offers in the UG has to be accounted to the very unfair nature of these offers, or to the fact that responders simply do not like to get less than proposers. In **chapter 4** I will present three behavioral studies in which I have manipulated both the allocator of the offers and their advantageousness: responders had to decide on offers made by either an external allocator or even a random number generator, which could be fair, unfair disadvantageous or unfair advantageous. Moreover, they were also asked to play on behalf of a third party, as in the previous two studies. Results showed that people tend to reject low offers, even if this does not mean to punish the source of unfairness; moreover, when playing for themselves, they are much more tolerant towards self-advantageous unfairness (or inequality), whereas, when playing on



behalf of a third party, they generally reject inequality. In **chapter 5** I report the imaging results obtained administering the modified UG described above to the participants. Results show a major activation in the MPFC for unfair offers in MS but not in TP, especially for disadvantageous offers; IFG/AI's activation was higher when facing unfair offers, both disadvantageous and advantageous, irrespectively of the target.

To conclude, rejections in the UG do not always correlate with factors that have been described as the cause of rejections (inequality/unfairness, negative reciprocity, and negative emotions). Models that take into account that preferences may vary with contextual cues, e.g. self-involvement, as described here, merit, personality traits, etc., are the best candidates to explain socio-economic behavior. As far as the neural correlates are concerned, I propose that IFG/AI signals the deviation from an expected or a desired outcome. To understand whether equality can be considered either an expected or a desired outcome, further work is needed.

# Chapter 1

## Introduction

*Decisions are always complex.*

Humans make hundreds of decisions in everyday life, adjusting their behaviour depending on the feedbacks they receive by the surrounding environment: ranging from whether having another coffee to whether having kids, all choices we make require a constant integration of many different factors. For this reason, the features characterizing a decision vary along with the particular situation in which a person is acting. For instance, imagine to be driving alone, on a deserted road, when you eventually reach a junction: the map is not very clear, but, as far as you can guess, either road could take you to your destination; however, while you are familiar with the road on your left, you have never taken the road on your right, which, from the map, seems to be the shortest. What are you going to do? Your decision depends on many factors: are you in a hurry? Are you a risk-taking person who likes novelties? Is the map reliable? This situation well exemplifies a condition in which the decision is made under uncertainty, given that not all the alternative outputs are known to you; moreover, any decision you take is going to influence only

yourself, as far as that particular situation is concerned (individual decision-making). Here is another example: you are driving the same old car, but this time you are together with a friend, when you find yourself to a junction. You are familiar with both the road on the right and the road on the left: the first one is faster, but with a lot of turns, and you know that your friend suffers from car sickness, and thus prefers the road on the left. However, you are a bit in a hurry, so what are you going to do? Again, your decision will depend on different factors, such as how much in a hurry you are and how much you care for your friend; this time you know precisely each of the outcomes and, also, the decision will influence others too, characterizing your decision-making as social. Despite the dichotomies between individual and social decision-making or between decision-making under certainty or uncertainty, all kinds of decisions share the complexity of integrating different environmental, cognitive and emotional factors. This complexity has led scholars to approach this issue from various perspectives and at different levels of analysis.

In the following paragraphs, I will briefly review the main perspectives within which decision-making has been addressed, focusing specifically on social decision-making, the main topic of this dissertation.

## **1.1 Different approaches to study decisions**

### *1.1.1 Game Theory*

Game Theory (GT) is a toolkit of mathematical models for analyzing the way in which decision-makers interact with and influence one another; it also offers a large amount of interactive situations that can be used as behavioral tasks to test its predictions. However, the viewpoint of classical GT is rational rather than psychological or sociological, as stated by the

Nobel Laureate Robert Aumann when defining the concept in The New Palgrave Dictionary of Economics; for this reason, the theory is grounded on assumptions, some of which may not exactly be satisfied in a typical real-world situation.

The first important assumption is that, in order to rank the set of outcomes to choose from, an individual can assign a value to each of the possible outcomes, which reflects her own preferences; these preferences are stable and consistent within each individual. Moreover, Game Theory assumes that players are rational, meaning that their objective is to maximize their own payoff, whose value is measured on a utility scale, that gives an index of the player preferences (see Utility Theory as described by von Neumann and Morgenstern, 1944). Game Theory also assumes that players are intelligent, which means that they know everything about the game and can infer everything as the game theorist does; this leads to consider players strategic, expecting them to choose the strategy resulted from the maximization of the expected utility.

As aforementioned, Game Theory describes a large amount of strategic situations, i.e. games, which according to Mas-Collel, Whinston, and Green (1995) can be described by four elements: players, who make decisions, rules, specifying what players can do and what they know, outcomes and preferences. These are *the matching pennies*, *the prisoners' dilemma*, *the battle of sexes*, just to mention a few of them (see e.g., Watson, 2008, for an extensive review). In the *prisoners' dilemma*, for instance, there are two players A and B (prisoners), who simultaneously choose to cooperate or to defect; the outcomes are all the possible combinations of defection and cooperation, i.e.  $\{(C,C), (D,C), (C,D), (D,D)\}$ , as expressed in the following payoff matrix.

B \ A	cooperate	defect
	cooperate	defect
cooperate	A and B both get 6 months	B gets 10 years A goes free
defect	A gets 10 years B goes free	A and B both get 5 years

Figure 1.1: Payoff matrix for the prisoners' dilemma.

The preferences on the utility scale are then  $DC > CC > DD > CD$ .

The solution concept of the game is a mathematical rule for generating predictions; the most important and well-known solution concept in Game Theory is the *Nash equilibrium*<sup>1</sup> (Nash, 1950), which represents a combination of strategies, one for each players, with the property that each player's strategy is optimal given each other player's choice. In the case of the *prisoners' dilemma*, Nash equilibrium is that both players choose to defect, the outcome (D,D): in fact, if both players cooperate they maximize both their payoffs, but neither player has an incentive to begin cooperating, given that unilateral cooperation would shift the player from the third worst payoff to the very worst one.

---

<sup>1</sup> Nash Equilibrium is defined as a set of strategies, one for each player, such that no player has incentive to unilaterally change her action. Players are in equilibrium if a change in strategies by any one of them would lead that player to earn less than if she remained with her current strategy. For games in which players randomize (mixed strategies), the *expected* or average payoff must be at least as large as that obtainable by any other strategy.

As I previously mentioned, Game Theory offers a bunch of descriptions of strategic and interactive situations, which can be used to test experimentally its predictions. However, empirical studies showed that standard economic analysis, which attributes to the decision maker both the will to maximize her own income and unbounded reasoning capabilities, failed to predict human behavior in many situations.

### *1.1.2 Behavioral Economics: social preferences and bounded rationality.*

As defined by Gul Faruk in the The New Palgrave Dictionary of Economics, behavioral economics refers to the research program that investigates the relationship between psychology and economic behavior; in fact, it has been seen that taking into account psychological variables, such as vengeance motives or fairness concerns, and human's limited cognitive capacity, is crucial in order to predict correctly human behavior. Many theories have been developed to incorporate these variables in formal socio-economical models; in particular, theories of social preferences try to provide a formal explanation for the apparently irrational behavior of people facing a certain kind of interactive situation. A more detailed description of these theories is given ahead in this chapter, when the specific case of the Ultimatum Game is taken into consideration.

Whereas theories of social preferences address the issue of the apparent lack of rationality by investigating humans' preferences for principles other than the maximization of monetary outcome, models of bounded rationality focus on people's cognitive constraints that limit the amount of available information. In 1957, in his *Models of Man*, Herbert Simon introduced the term *bounded rationality* in opposition to *full rationality*, *maximization of expected utility*, or simply *optimization*. The question he asked himself was: "how do human

beings reason when the conditions for rationality postulated by the model of neoclassical economics theory are not met?” Simon suggests that economic agents employ the use of heuristics to make decisions rather than a strict rigid rule of optimization, in order to face together the complexity of the situation, and their inability to process and compute the expected utility of every alternative action. The 2002 Nobel Laureate Daniel Kahneman distinguishes “two modes of thinking and deciding, which correspond roughly to the everyday concepts of reasoning and intuition” (Kahneman’s Prize Lecture, December 8<sup>th</sup>, 2002); he claims that reasoning is at work when we deliberately –and effortfully- compute something, such as a mathematical problem or the road on a map, while intuition is what makes us “reluctant to eat a piece of what we know to be chocolate that has been formed in the shape of a cockroach” (Kahneman’s Prize Lecture), and since it comes spontaneously to mind, without computation, it is also the most exploited system of the two. A different approach to bounded rationality is the one put forward by Gerd Gigerenzer and collaborators (Gigerenzer et al., 1999), who demonstrated that individuals and organizations often rely on simple heuristics which are better off than formal logic to solve problems of everyday life, given that ignoring aspects of the information can lead to more accurate judgments than weighting and adding all different types of information, like for instance for low predictability events and small samples.

### *1.1.3 The contribution of Cognitive and Brain Science and the rise of Neuroeconomics.*

The same issues about how people make decisions have been addressed also by psychology and cognitive sciences, that describe these phenomena using a less formal approach, as it is typical of these disciplines, and introducing other concepts such as personality traits, emotions and moral sentiments. Cognitive sciences aim at investigating the influence of

psychological factors on the behavior, such as the correlation between anger and frustration with decisions in bargaining games (see the wounded pride/ spite model by Pillutla and Murnighan (1996), discussed more in details in the next paragraph). However, social and psychological sciences can, at most, infer the cognitive and emotional underpins, whereas the neuroscientific methods that have been recently developed can allow the identification of the micro-foundations of the cognitive activity in the neural system.

For instance, the feedback from the environment is crucial in order to successfully adjust our decisions; such adjustment depends on the ability to discriminate between negative and positive feedbacks, which indicate the appropriateness of behaviour (Nieuwenhuis, Holroyd, Mol, and Coles, 2004). This important aspect of decision-making has been extensively studied in recent years. The method that better mirrors feedback processing is the detection of the event-related potentials (ERPs), which have been largely employed in these studies. Another example is the involvement of emotions in decision-making, an issue that has been recently given a lot of attention: thanks to psychophysiological methods, such as skin conductance response (SCR) and neuroimaging, especially functional magnetic resonance (fMRI), it has been possible to investigate this issue in much more details, obtaining quantitative data to correlate to the subjective self-report psychological scales. In discussing the role played by emotions in decision processes, it is essential to describe a very influential view, which is called *the somatic marker hypothesis* (Bechara & Damasio, 2005). The somatic marker hypothesis posits that emotion-related bodily signals assist cognitive decision-making, supporting the idea that knowledge and reasoning alone are usually not sufficient to make advantageous decision. This concept is in contrast with the traditional point of view whereby emotion is considered to perturb rational decisions. In contrast, in Bechara and Damasio's model, emotions can be either beneficial or



disruptive depending on whether they are integrated into the task or unrelated to it, respectively. Emotion is defined as a collection of changes in body and brain states triggered by brain structures such as amygdala, insula, ventro-medial prefrontal cortex and the brainstem. Several empirical evidences support the somatic marker hypothesis, and most of the studies employed the Iowa Gambling Task, the skin conductance response and brain damaged patients (Bechara A., Damasio, Tranel, and Damasio, 2005; Damasio A., 2005; Bechara A., Damasio, Tranel, and Damasio, 1997; Bechara A., Damasio, Damasio, and Lee, 1999). Even though many criticisms have been moved to this theory (see, for example, Tomb, Hauser, Deldin, and Caramazza, 2002), it is undoubted that the contribution of emotions to decision making processes is a relevant issue to explore.

The emerging neuroeconomic approach seeks a microfoundation of social and economic activity in neural circuitry, using different methods such as functional magnetic resonance imaging (fMRI), transcranial magnetic stimulation (TMS), pharmacological interventions and other techniques. The neuroeconomic approach aims at unifying mechanistic, mathematical and behavioral measures and constructs, in order to better understand:

*“individual differences and development over the human lifecycle (including disorders and expertise), insights into the effects of direct and social learning, empirical discipline of evolutionary modeling, and advice for how economic rules and institutions can be designed so that people react to rules in a socially efficient way.” (Fehr & Camerer, C., 2007).*

Given that the work described in the next chapters is focused on one single game, specifically the Ultimatum Game, I will now discuss this issue more in details; in order to clarify the theoretical frame shaped on this specific, I will briefly review the previous findings in literature, and focus on the open questions that have provided the starting point of my research.

## 1.2 The Ultimatum Game

The Ultimatum Game (UG) is one of the classical paradigms used to investigate economic decision-making, and is taken as paradigmatic evidence to illustrate the failure of classical economic theories in predicting human behavior. In its traditional version, developed by Güth and colleagues about thirty years ago (Güth, Schmittberger, and Schwarze, 1982), one player (the proposer) makes offers to a second player (the responder) on how to split an amount of money given by the experimenter; the responder, in turn, can either accept or reject the offer. If the responder accepts, the money will be divided as suggested by the proposer; if the responder declines the offer, both players end up with nothing.

Classical economic theories (e.g., Von Neumann and Morgenstern, 1944) posit that the proposer, in order to maximize the gain, should always offer the smallest amount of money, whilst the responder, following the principle that “few is better than nothing”, should always accept the offer; this constitutes the Nash equilibrium for this game. In contrast with this prediction, behavioral findings clearly show that the proposer tends to divide the money equally, and that the responder rejects unfair offers which favor the proposer too much. Importantly, this behavioral pattern has also been observed in both the “single-shot” version of the UG, in which the two players interact once only, and in the “covered” version of the UG, in which the proposer is not informed about the responder’s reaction (Abbink, Sadrieh, and Zamir 1999; Zamir, 2001; Civali, Corradi Dell'Acqua, Gamer, and Rumiati 2010), making rejections losing their negotiating role.

### *1.2.1 Theories of social preferences.*

Why do people behave against their self-interest? Two theoretical approaches have been developed in order to explain the apparently irrational behavior in the UG. The first one claims that rejections are grounded in the strategic thinking: *proposers* offer more than what is predicted by game equilibrium because they know that otherwise their offers will be rejected, increasing strategically their chances to maximizing their expected payoffs (Weg and Zwick, 1994). The second approach takes into account the so-called social preferences: this account posits that people, being endowed with some sense of fairness, may not only care about their self-interest, but also about the interest of others (e.g., Guth, 1988). To disentangle between these two approaches, a crucial variant of the UG, called the Dictator Game (DG), was introduced by Kahneman, Knetsch and Thaler (1986): in this game, the proposer becomes a dictator, since the offer cannot be rejected by the responder, who, therefore, becomes a simple receiver. Results show that a considerable percentage of people are driven by a taste for fairness, offering more than the minimum offer predicted by classical utility theories, although offers are lower than in the UG, and there are exceptions that will be discussed in the following sections. These evidences suggest that, even though, to a certain extent, strategic thinking affects offers (offers in DG are lower than in UG), fairness concerns may play an important role. Within the social preferences theoretical framework, several models have been developed in the last years (for an overview see Camerer, 2003). Theories of *inequity aversion* (Bolton, 1991; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) claim that an agent's utility is completely determined by the final distribution of outcomes, and that people are motivated by a general distaste for unequal outcome: player B will always reject an unfair distribution, even though it is decided by the fortune. By contrast, theories of *reciprocal fairness* or *negative reciprocity* (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006) assume that “people like to

help those who are helping them and to hurt those who are hurting them” (Rabin, 1993). These models take into account the influence of the actual actions and beliefs on agent's utility in determining the decision: if player A reduces the payoff of player B in order to obtain benefits, player B will negatively reciprocate player A by punishing his/her bad intentions, whereas if incomes' distribution is decided by the fortune, player B will not punish player A for getting a higher payoff (Blount, 1995; Charness, 1996; Falk & Fischbacher, 2006). In this perspective, rejections in the UG can be interpreted as a tool in the hand of the responder to punish a proposer who behaved unfairly.

In order to study these issues, I have run a series of three experiments, described in chapter 4, which aimed at shedding light on the role played both by inequity aversion and by negative reciprocity on UG's rejections.

In a psychological perspective, negative emotions, such as frustration, have been elected as the ultimate cause of rejections: Pillutla and Murnighan (1996) described the wounded pride/spite model, which claim that perception of unfairness can lead to anger and wounded pride and, ultimately, to a spiteful rejection. Some of the main neuroscientific findings that, thanks to the innovative approach, helped to shed light on the psychological basis of the UG behavior will be reviewed below.

### *1.2.2 Psychophysiological studies: skin conductance response (SCR)*

When looking at the effective presence of emotional arousal, a straight and relatively simple way to detect it is to measure the electrodermal activation. This represents the level of activation of the sympathetic system, detected by placing two active electrodes on two fingers of

one hand and applying a small amount of current; the conductance (or the resistance) of the skin is represented by the quantity of sweat present in the sweating ducts: the more the amount of the sweat, the higher is the conductance (or, conversely, the smaller is the resistance). The receptors present in the hands' palms refer are known to be specifically sensitive to cognitive and emotional efforts more than to the temperature regulation; for this reason, this index is particularly suitable to detect the effects of the cognitive and especially emotional arousal (Boucsein, 1992). One disadvantage of this technique is that it is a-specific: it cannot disentangle among the different types of arousals, and, therefore, among the different kinds of emotions. To overcome this problem, it is necessary to ask for self-reported impressions to the participants – for example, by administering emotional rating scales-, in order to match the physiological index with the effective cognitive-emotional state.

The emotional model put forward by Pillutla and Mourninghan has been supported by a study carried out by van't Wout and colleagues (van't Wout, Kahn, Sanfey and Aleman, 2006): the authors found that participants playing as responders in a classical UG showed a higher SCR when rejecting, as opposed to accepting, unfair offers made by a human being, whereas in the control condition, in which they were playing against a computer, they did not show this difference between responses. They concluded that, negative emotions trigger rejections. However, the next chapter will discuss a recently published study, in which we have shown that rejection of unfairness is not necessarily correlated to emotional reactions.

### *1.2.3 Brain imaging studies: functional magnetic resonance (fMRI)*

The use of imaging techniques has spread widely in recent years, and the majority of the research centers nowadays have the availability to a scanner for magnetic resonance

measurement. In particular, the more popular is the functional magnetic resonance, which offers the possibility to get a functional image of the brain while the participant is administered with a task. The cerebral activity is measured as the blood oxygen level-dependent (BOLD) signal, which represents the level of metabolism in a specific area of the brain: the assumption underneath is that the more an area is active, the more is the oxygen required to that area, and the stronger the signal. This technique has some limitations, such as the poor temporal resolution, as compared to other techniques such as the electro-encephalography, and the fact that the signal actually represents an indirect measure of the cerebral activity; however, it is non-invasive, safe and relatively available, thus the most popular imaging technique.

Sanfey et al. carried out an fMRI study to investigate which brain areas were activated in participants playing as responders, finding that dorsolateral prefrontal cortex, anterior cingulate cortex and anterior insula were more active when processing unfair as opposed to fair offers. In particular, the anterior insula, an area traditionally associated to negative emotions such as disgust (e.g. Calder, Lawrence and Young, 2001), was more active during rejections as opposed to acceptances; the authors interpreted the results as a proof of an involvement of negative emotions in triggering rejections (Sanfey, Rilling, Aronson, Nystrom and Cohen, 2003). While Sanfey and colleagues focused on areas involved in unfairness and rejections, Tabibnia and colleagues paid their attention to the areas associated with the perception of fairness: they found that commonly identified reward areas, such as ventral striatum, amygdala and ventromedial prefrontal cortex, were associated with the preference for fair outcome (Tabibnia, Satpute, and Lieberman, 2008). Another interesting result concerning the reward system, which is the brain circuit activated by a reward administered to the subject, comes from the positron emission tomography (PET) study of de Quervain et al. (de Quervain, et al., 2004): the results showed an

activation in the dorsal striatum, part of the reward system, when subjects were actively punishing unfair individuals, supporting the idea that punishing violators of social and moral norms is satisfactory.

In chapters 3 and 5 I will report two fMRI studies in which the link among unfairness, rejections and crucial brain areas such as anterior insula and medial prefrontal cortex is further clarified.

#### *1.2.4 Neuropsychological studies: prefrontal patients*

The study of brain damaged patients is one of the oldest methods to investigate the connections between brain and behavior. As already mentioned when talking about the Somatic-marker hypothesis, patients with a lesion of the ventromedial part of the prefrontal cortex (vmPFC) show deficits, especially as far as decision-making is concerned.

Koenigs and Tranel (2007) investigated the responder behavior of ventromedial prefrontal patients and showed that these patients were more prone to frustration, since they were rejecting more unfair offers compared to controls, suggesting that this area is involved in the emotional control.

Partially in contrast with these results, Moretti, Dragone, and di Pellegrino (2009) found that vmPFC patients actually rejected more unfair offers, as opposed to controls, but only when the financial gains were presented as abstract amounts, whereas, when they were immediately delivered to the patients, their rejection rate showed no difference from controls; these results suggested that vmPFC is involved in processing the expected values of abstract and future goals more than the feeling of frustration associated to the actual social value of the offer.

### *1.2.5 Brain stimulation studies: transcranial magnetic stimulation (TMS), transcranial direct current stimulation (tDCS), tryptophan depletion.*

So far, I have described methods that show correlations between cerebral or physiological activity, but that, apart from patients' studies, do not allow any inference about causality. I am now going to review some studies in which brain stimulation techniques were used, allowing some causal hypothesis between brain areas or physiological processes and behavior in the UG.

Knoch and colleagues (Knoch, Pascual-Leone, Meyer, Treyer and Fehr. 2006) applied the repetitive transcranial magnetic stimulation (TMS) to the right dorsolateral prefrontal cortex (dlPFC) of participants playing as responders. This technique consists in administering subjects with a magnetic impulse on a specific brain area, which causes a perturbation in the regular electrical neuronal activity, temporarily disrupting the functionality of that particular area; for this reason, it is also called "virtual lesion technique". The results showed that the disruption of the activity in the dlPFC caused a decrease in the rejection rate of participants, despite the offers were still considered unfair: following the authors' interpretation, this suggests that this area plays a crucial role in fairness perception, being important to override self-interest in order to implement fairness goals. In a further study, Knoch et al. (Knoch, et al., 2008) replicated the results using another stimulation technique, the transcranial direct current stimulation (tDCS): in this case, not a magnetic impulse, but a direct electric current interfere with the neuronal activity of the area underneath the electrodes.

Another interesting and promising avenue of research is the pharmacological approach, which consists in interfering with the brain activity by modifying the metabolism of a particular neurotransmitter. In a recent work, Crockett and colleagues (Crockett, Clark, Tabibnia, Lieberman, and Robbins, 2008) showed how decreasing the amount of serotonin, by the



depletion of its precursor amino acid (tryptophan) causes an increase in the UG rejection rate, supporting a correlation between the emotional effects of the lack of serotonin and the reaction to unfairness.

#### *1.2.6 What is there in a rejection? Open questions and hypothesis.*

So far, all the evidences seem to converge towards a strong involvement of emotions in the rejections of unfairness: the correlation between unfairness and frustration which leads, consequently, to an irrational rejection seems to be straightforward. However the UG is, for its own definition, a self-centered task: the offers address only the responder, and so does the potential unfairness. Therefore, it is impossible to say whether the frustration is elicited by the pure perception of unfairness or by the fact that the unfairness is damaging the responder herself. The modified version of the UG I have developed in order to disentangle between these two possible causes is described in the next chapter, and it has been used to test both the emotional arousal (chapter 2) and the neural activation (chapter 3).

Moreover, in the traditional version of the UG, responders face offers which are unfair but also disadvantageous for her, making unclear whether responders' aversion towards low offers in the UG has to be accounted to the very unfair nature of these offers, therefore supporting an endowed preference for fair outcomes, or to the fact that responders simply do not like to get less than proposers, focusing on a more selfish motive. Finally, while it is true that more than one study has proven that intentions of the counterpart matter when deciding whether to accept or reject an allocation, however they cannot account *in toto* for the deviation from classic rationality. I have addressed these issues by further manipulating the traditional UG, and chapters four and five report behavioral and imaging results.

## **Chapter 2**

# **Are irrational reactions to unfairness truly emotionally-driven? Dissociated behavioral and emotional responses.**

### **2.1 Introduction**

In recent years the study of the role of emotions in decision-making has become an increasingly prominent issue in cognitive neuroscience. A wealth of studies have hypothesized an emotional pathway in the brain that seems to operate in many types of decisional processes, including moral judgment (Moll, Zahn, de Oliveira-Souza, Krueger, & Grafman, 2005, for a review) and economic decision-making (Pillutla & Murnighan, 1996; Sanfey et al., 2003), that have traditionally been linked to rational thinking and choices (Kohlberg, 1969; von Neumann & Morgenstern, 1944).

The Ultimatum Game (UG) has always been thought of as a classical example of emotionally-driven behavior. It has been argued that self-centered emotions, such as anger and

frustration, play a crucial role in the UG, as the individual payoff is heavily involved (Moll & de Oliveira-Souza, 2007). However, it has also been proved that individuals choose to punish unfairness even though the violation of fairness and cooperation norms does not affect directly their payoffs (i.e. the third-party punishment): Fehr and Fischbacher (2004) found that participants decided to give up some of their own money to punish the unfair behavior of one player towards another. Thus, altruistic punishment, which is an act of punishment that, even if costly and yielding no direct benefit for the punisher (as in the case of single-shot UG), is used to penalize selfish behavior of others, as it leads them to cooperate in future interactions (Fehr & Gächter, 2002), also occurs in conditions in which unfairness should not elicit, at least in principle, any self-centered emotion. This raises the question of whether, in the UG task, the “irrational” punishing behavior and negative emotions are always causally related, or whether they can operate separately depending on the myself / third-party distinction.

In the present study, I have investigated the role of emotions in the UG by measuring skin conductance responses (SCR) while participants played as responders in a modified version of the UG and by collecting emotional ratings after they completed the task in order to measure the valence of the hypothetical arousal. Participants carried out both the classical version of the UG and a modified version of the task in which any putative monetary income was not going into the participants’ own pocket, but into a third-party’s (see Method section). Indeed, as in the latter condition the proposer’s offer did not directly address the participant’s payoff, unfairness should, at least in principle, elicit neither self-centered emotions (Moll & de Oliveira-Souza, 2007), nor, as consequence, SCR increases when the offer is about to be rejected (van’t Wout et al., 2006). Thus, the account according to which the punishing behavior and the negative emotions are causally related (Fehr & Gächter, 2002; Pillutla & Murnighan, 1996), also predicts that such

emotional decrease should be associated with a similar decrease in the amount of punishing choices (rejections) (van't Wout et al., 2006). However, based on previous studies on altruistic punishment (Fehr & Fischbacher, 2004), I predicted that participants should reject unfair offers addressing a third-party; if this were indeed the case, a significant increase in SCR was expected, for offers about to be rejected even in the third-party condition.

## **2.2 Methods**

### *2.2.1 Participants*

Thirty-four healthy Italian volunteers (22 females), who ranged in age from 18 to 35 years ( $M=23.56$ ,  $SD=3.90$ ), took part in the experiment. They all were paid for participating in the study, the scientific goal of which was unknown to them. The study was approved by the local ethics committee and conducted in accordance with the Declaration of Helsinki.

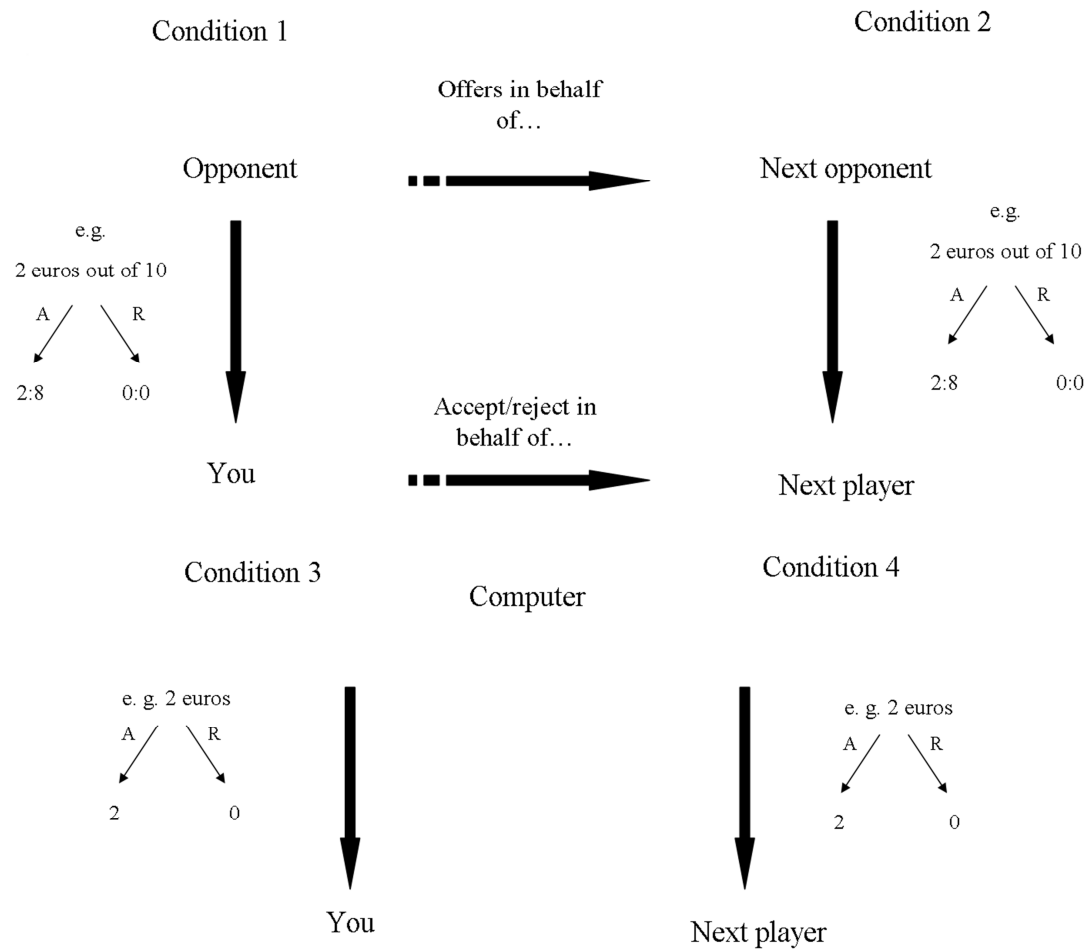
### *2.2.2 Task*

Participants were required to play as responders in a modified version of the UG and had either to accept or reject the offers the proposer made, following the classical rules explained above. Before starting the game, they were introduced to a collaborator of the experimenter, who pretended to play as the proposer, in order to strengthen the illusion of playing against a human adversary, whereas they were actually playing against a computer. They were told that the opponent had been given a number of 10 euros bank notes and would have to make offers on how to split each of them. Consistent with previous studies (Polezzi, et al., 2008), offers in each trial could be either 1, 2, 3, 4 or 5 euros out of 10. Furthermore, participants were informed that, in one condition, they and their opponent would play for themselves (consistent with the

classical UG), whereas, in another condition, they would play on behalf of those players acting as proposer and responder in the upcoming testing session (see Figure 2.1). In order to make our task compatible to the single-shot UG, participants were told that the opponent would receive feedback only at the end of the experiment, when they have both been informed on how much each of them had gained, depending on the choices they had made; in this way, they knew rationally that they could not affect the opponent's behavior through their rejections.

To control for the social interactive nature of the UG, participants performed a control task (Free Win [FW] task) in which they either accepted or rejected a variable amount of money given by the computer (1, 2, 3, 4 or 5 euros). As in the case of the UG, they could decide for themselves or on behalf of the next participant. If they accepted the offer, they/the third party would receive that amount; otherwise they/the third party would receive nothing. This yielded to a 2 x 2 x 5 design, with TASK (UG vs. FW), TARGET (myself -MS- vs. third-party -TP-) and GAIN (1, 2, 3, 4 or 5 euros) as within-subjects factors.

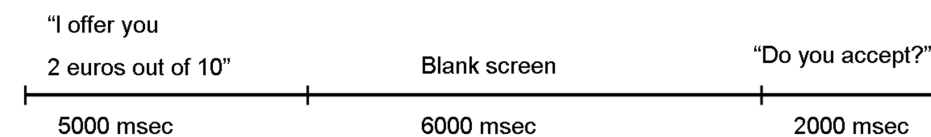
Participants were informed that their compensation for participating in the experiment would be proportional to the amount of money gained in the MS condition. Moreover, they knew that a percentage of the money split on behalf of third parties would be given to next players; they were also informed that, following the same principle, their starting stakes were percentages of the money that previous players had split on their behalf. Irrespective of their performance on the task, participants received the same amount of money as compensation. Although we did not systematically investigate whether participants had doubts about the authenticity of the situation, the majority of them, when informally interviewed afterwards, said they believed they had played against a human opponent. Only a few reported having doubts at the end of the experimental session.



*Figure 2.1.* Illustration of the task as it was presented to participants when giving the instructions. There are four conditions: the first and the second refer to the Ultimatum Game and the third and the fourth refer to the control task (Free Win situation). In the first and in the third conditions participants are asked to decide for themselves, whereas in the second and in the fourth they are asked to choose on behalf of a third party (next participant).

### 2.2.3 Apparatus and Procedure

All the participants were tested in a quiet room at SISSA using a pc and a 15-in monitor (Olidata s.p.a.). Presentation® 12.0 software (<http://www.neurobs.com>) was used to construct and deliver the experimental stimuli. The offer appeared on the screen for five seconds, followed by a six-second blank screen. Participants were required to respond by button press, highlighted on the computer keyboard, as soon as the question “Do you accept?” appeared on the screen, where it lasted for two seconds (see Figure 2). The inter-trial interval was around 11 seconds on average, to allow skin conductance to return to its baseline. All 20 conditions, each of which was repeated four times, were presented in a randomized order. The whole experiment (80 trials \* 24 seconds of trial duration) including a short break of one minute after half of the trials lasted approximately 33 minutes.



+ ITI avg 11000 msec

*Figure 2.2.* Time line for each single trial of the UG. Each trial lasted 24 seconds. First, participants saw the offer on the screen for five seconds, followed by six seconds of blank screen. Next, the question “do you accept/do you accept on his behalf” appeared on the screen for two seconds, within which participants had to answer by button press. An average of eleven seconds inter-trial interval followed the question.

### 2.2.4 Skin Conductance Recordings

Skin conductance was recorded during the whole experiment using a pair of prewired 8 mm Ag/AgCl electrodes, attached to the distal phalanx surfaces of the index and little finger of the non-dominant hand. The electrode pair was excited with a constant voltage of 0.5 V and conductance was recorded using a DC amplifier with a low-pass filter set at 64 Hz and a sample frequency of 256 Hz. Values of skin conductance were automatically transformed to microsiemens values by the Procomp Infinity System (Bio-Medical Instruments, Inc., Warren, MI, USA). Before starting the task, one minute of baseline was recorded. I measured the artifact-free amplitude of the skin-conductance response that began between 1 and 3 seconds after the presentation of the offer and exceeded a threshold of 0.05  $\mu$ S. In the case of overlapping responses, the inflection point between the two responses served as the baseline or peak, depending on the latency criterion. The resulting amplitudes were z-transformed within each participant in order to eliminate individual differences in responsivity.

#### *2.2.5 Emotional Ratings*

To further investigate the emotional reactions in this study, participants rated their feelings in the most crucial conditions (i.e., 1, 3 and 5 euros of gain when playing the UG in both the MS and the TP condition) at the end of the experimental session. To this aim, they used a 12-points Likert-scale for each condition ranging from -6, corresponding to strong negative emotions, to +6, indicating strong positive emotions.

## **2.3 Results**

### *2.3.1 Rejection Rates*



For each subject and condition, the rejection rates were calculated across all 4 repetitions, and used in a 2 (TASK: UG, FW) x 2 (TARGET: MS,TP) x 5 (GAIN: 1, 2, 3, 4, 5 Euros) Repeated Measures ANOVA. Statistical Analysis was carried out using SPSS 11.5 Software (SPSS Inc., Chertsey UK). Results indicated a significant main effect of TASK ( $F(1, 33) = 76.24, p < .001, \eta_p^2 = .69$ ), with the UG eliciting a larger amount of rejections than the FW (see Table 2.1 and Figure 2.3), as well as a main effect of GAIN ( $F(4, 132) = 52.7, p < .001, \eta_p^2 = .61$ ), with low offers being rejected more than high offers. This effect is however driven by the TASK x GAIN interaction, which was found to be significant as well ( $F(4, 132) = 49.89, p < .001, \eta_p^2 = .60$ ), suggesting that low offers are rejected significantly more often than high offers in the UG but not in the FW. None of the remaining effects of the ANOVA were found to be statistically significant.

*Table 2.1* Rejection rates (RR) (percentages) and skin conductance response amplitudes (SCR amp) (z-transformed  $\mu S$ ) for the four conditions collapsed by gain.

	UG		FW
	Myself	Third-party	Myself
RR (SEM)	35.73 (5.59)	38.09 (4.93)	2.64 (0.68)
SCR amp (SEM)	.0887 (.0157)	-.0073 (.0164)	.0257 (.0146)

*Note.* Corresponding standard errors of the mean are printed in brackets.

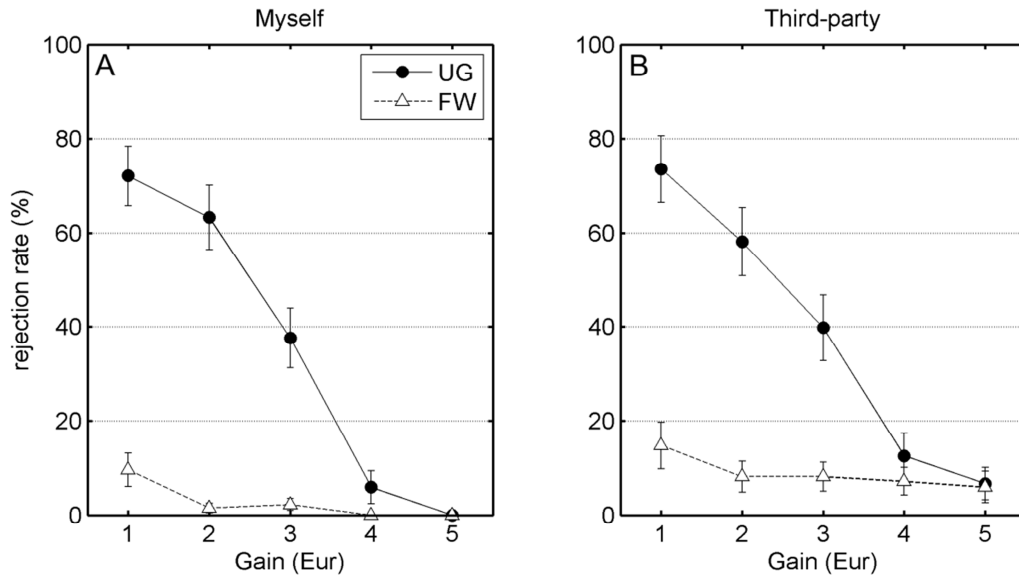


Figure 2.3. Behavioral results. Rejection rates (in percent) plotted as a function of GAIN in the MA (A) and the TP (B) condition.

### 2.3.2 Emotional Ratings

The emotional ratings for the most unfair offer (1 euros out of 10), the fairest offer (5 euros out of 10) and the mid-value offer (3 euros out of 10) were analyzed, both for MS and TP conditions. One-sample two-tailed T-tests showed that for the mid-value offer the ratings did not differ significantly from zero (i.e. the neutral emotion), while for both targets, the ratings for the unfair offer were significantly different from 0 towards the negative emotion (UG (1:9) MS:  $t(33) = -9.79, p < .001$  UG (1:9) TP:  $t(33) = -4.37, p < .005$ ), and so were those for the fair offer towards the positive emotion (UG (5:5) MS:  $t(33) = 22.29, p < .001$ ; UG (5:5) TP:  $t(33) = 5.63, p < .005$ ). Moreover, an ANOVA, considering TARGET (MS and TP) and GAIN (1, 3, 5 euros out of 10) as factors, showed a significant effect of TARGET ( $F(1,33)=4.328, p < .05, \eta_p^2=.116$ ), a significant effect of GAIN ( $F(2,66)=101.82, p < .001, \eta_p^2=.75$ ), and a significant TARGET x

GAIN interaction ( $F(2,66)=12.662$ ,  $p<.001$ ,  $\eta_p^2=.277$ ). A paired-samples T-test demonstrated that there was a significant difference between targets for fair ( $t(33)=4.01$ ,  $p<.001$ ) and unfair ( $t(33)=-2.742$ ,  $p<.01$ ) offers, while no difference between targets were found for the mid-value offer; in particular, both the reported positive and the negative emotions were rated as stronger in the MS than in the TP condition.

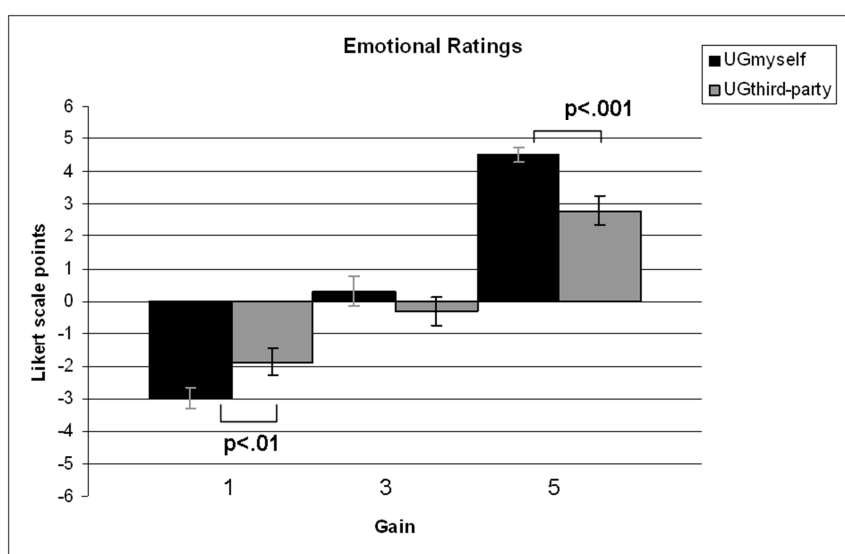


Figure 2.4. Emotional ratings. The black bars indicate the MS condition, while the grey bars indicate the TP condition, for gain 1, 3 and 5. Error bars indicate the standard deviation.

### 2.3.3 Skin Conductance Response amplitude

For each subject and condition, the average of z-standardized skin conductance response amplitudes were calculated across all 4 repetitions, and used in a 2 (TASK) x 2 (TARGET) x 5 (GAIN) Repeated Measures ANOVA. A significant main effect of TASK ( $MSE = 0.23$ ,  $F(1,33) = 4.91$ ,  $p < .05$ ,  $\eta_p^2 = .13$ ) and a significant main effect of TARGET ( $MSE = 0.28$ ,  $F(1,33) = 7.93$ ,  $p < .01$ ,  $\eta_p^2 = .19$ ) were found, suggesting that participants were more aroused whilst

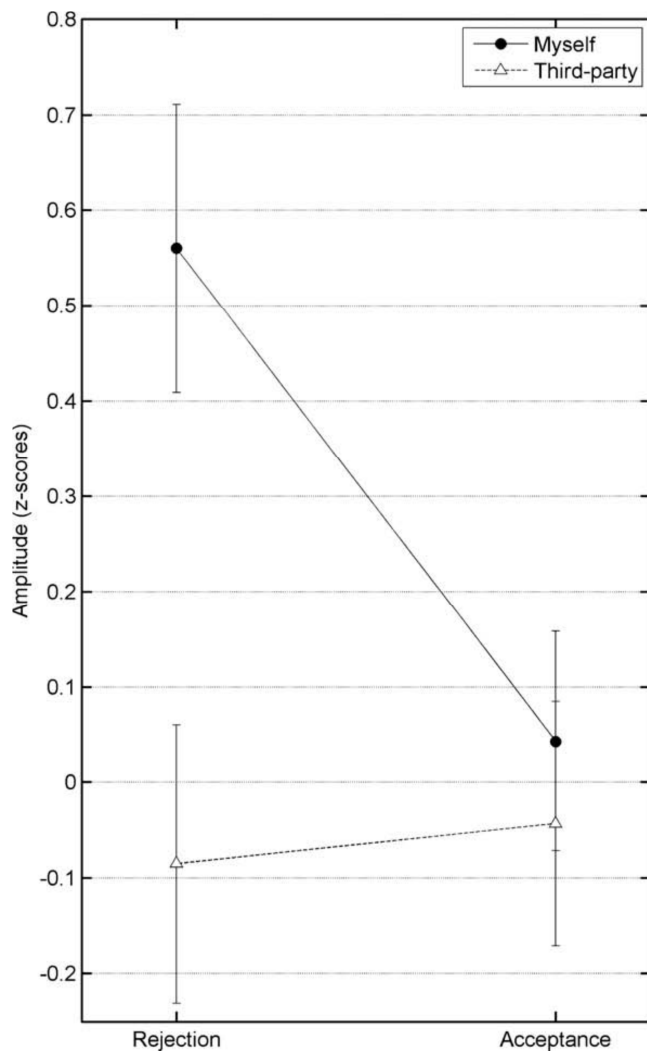
playing the UG than the FW, and when their own interest, and not the third-party's, was at stake (see Table 2.1). None of the remaining effects were found to be significant.

In addition, I investigated the relation between SCR and rejections. Following van't Wout et al. (2006), we focused our analysis on small offers (1 Euro), as in the UG they were associated with the largest negative emotional arousal (see our analysis of emotional ratings above). A Linear Mixed Model (Neuhaus, McCulloch, & Shayle, 2008) was used, which is more robust against missing cells, as few subjects scored in all conditions. The model included RESPONSE (accept/reject), TARGET and TASK as fixed factors, and SUBJECTS as random factor. A compound symmetry covariance structure was specified. I found a significant main effect of TARGET ( $F(1, 138.67) = 7.36, p < .01$ ), indicating a stronger emotional arousal when offers were directed to oneself ( $0.23 \pm .11$  z-transformed SCR) rather than to a TP ( $-3.46 \cdot 10^{-5} \pm .09$  z-transformed SCR), and a significant TARGET x RESPONSE interaction ( $F(1, 144.91) = 4.28, p < .05$ ), reflecting participants' higher SCR amplitudes when rejecting small offers for themselves than when rejecting for a third-party. No target difference was found for the acceptances (see Table 2.2 and Figure 2.5).

*Table 2.2* Skin conductance response amplitudes (z-transformed) for Rejections and Acceptances both for the Myself and for the Third-party condition.

	Myself	Third-party
Rejections	.560 (.151)	-.085 (.146)
Acceptances	.043 (.115)	-.043 (.128)

*Note.* Corresponding standard errors of the mean are printed in brackets.



*Figure 2.5.* Physiological results. Z-standardized skin conductance response amplitudes plotted as a function of RESPONSE for Gain 1. Full lines and filled circles refer to the MS condition, whereas dashed lines and empty triangles refer to TP condition. Error bars indicate standard errors of the mean.

## 2.4 Discussion

The nature of “irrational” rejections in the Ultimatum Game have been investigated, by having participants perform a modified version of the paradigm in which they were asked to play for themselves or on behalf of a third party. To this purpose, I considered the rejection rate of the

offers as a behavioral measure and both the related skin conductance activity and the subjective ratings as indexes of emotional activation. A dissociation between behavioral and emotional responses was found: participants rejected an equal amount of small offers in the UG (but not in the control task) irrespective of whether these addressed oneself or a third-party; however they exhibited an increased negative emotional arousal when about to reject the most unfair offer addressing oneself (but not a third party). The account according to which rejections in the UG are irrational responses driven exclusively by negative emotions should therefore be reconsidered.

The well-documented pattern of accepting fair offers and increasing the rate of rejection as offers become less fair was replicated (Bolton & Zwick, 1995; Roth, 1995; Guth, Huck & Muller, 2001; Sanfey et al., 2003). In keeping with what predicted by Fehr and Fischbacher (2004), participants showed the same behavior even when playing on behalf of a third party. This pattern was not found in the control task, in which participants had to either accept or reject money given by the computer. This allows to concluding that, even though the responder's personal gain is the same in both UG and control task, in the UG, only the perception of an unfair split of money drives him/her to reject these offers choosing the so called non-utilitarian or “irrational” solution.

The analysis of the electrodermal activity revealed an increase of offer-related SCR amplitudes whilst playing the UG, relative to the FW, and when one's own interest, relative to a third party's, was at stake. No significant effects associated with the factor GAIN were found thus suggesting an equal amount of emotional arousal irrespective of the magnitude of the offers. However, the analysis of emotional ratings revealed a significant increase of negative emotions associated with the most unfair condition (1 Euro out of 10), a significant increase of positive

emotions associated with the most fair condition (5 Euros out of 10) and no significant emotional activation during mid-value offers (3 Euros out of 10). Thus, the significant increase of SCR amplitudes is assumed to be associated with the both fair and unfair conditions in the UG (as opposed of the FW) could well reflect an increased emotional arousal irrespective of valence. In the case of the mid-value offers, it can be argued that SCR is more likely to reflect cognitive effort (Boucsein, 1992), as mid-value offers in the UG are usually associated with the longest response times, and with a larger N350 after the presentation of the offer (Polezzi et al., 2008), which usually occurs when ambiguous stimuli are processed (e.g. Schendan & Kutas, 2003).

Finally, when the focus was on the trials associated with small offers (1 Euro), which in the case of the UG are the most unfair and are associated with largest negative emotional activation, a significant increase of SCR was found when rejecting (rather than accepting) offers addressing oneself. Such an increase (reminiscent to the one first described by van't Wout et al., 2006) was not found when the offers were directed at a third-party. Thus, if rejections are emotionally driven, as they are not utilitarian in nature (Fehr & Gächter, 2002; Pillutla & Murnighan, 1996), we would expect to find an increase in the electrodermal activity when participants reject (compared with when they accept) small offers also when these address a third-party. Instead, these data suggest that participants' rejections and their emotional reaction are independent, although co-occurring when participants play the UG for themselves.

An alternative explanation for the responder's behavior can be related to the notion of context-dependent fairness proposed by Zamir and colleagues (Winter & Zamir, 2005; Zamir, 2001), according to which the sense of equity may change depending on both the person engaged in the social interaction dynamic, and the nature of this dynamic. For instance, Winter and Zamir (2005), reported a modified version of the UG in which the proposer played with virtual-

responders which could be either much more tolerant or unforgiving to unfair offers than real human responders. They found that the proposers quickly adapted their behavior to the virtual-responders, by behaving unfairly with the tolerant and fairly with the unforgiving responders. This is similar to what happens in the Dictator Game (Forsythe, Horowitz, Savin, & Sefton, 1994; Bolton & Zwick, 1995), in which the proposer cannot have his offers rejected by the responder and, therefore, behaves far less fairly than in the UG. All these observations suggest that, in the UG, proposers' behavior is directly affected by the tolerance to unfairness he expects in the responder. Even though rejections in the UG are irrational from an individualistic perspective, in that the money loss does not increase the responder's chance of having better offers in the remaining part of the experimental session, they can be considered rational from a collectivistic point of view, because they are supposed to lead the proposers to play fairly and, consequently, to an increase in the overall gain for the population of the responders (Zamir, 2001). The account according to which the responder's rejections are utilitarian is in agreement with behavioral results I have presented here. In this study, participants were told prior to the experiment that their starting stakes depended on how previous players had decided to split the money; it is therefore likely that they felt part of a group in which cooperation led to a maximization of everyone's gain. Thus the participants' rejections, on behalf of the third party, of the offers which are considered unfair, might reflect the will of preventing a bargain which, if accepted, would be detrimental for the population of the responders (Zamir, 2001). Critically, this account does not necessarily predict that the rejection should be associated with an increased negative emotional arousal.

That emotions do play a role in the UG is demonstrated by other studies (e.g., Harlé & Sanfey, 2007; Sanfey et al., 2003; van't Wout et al., 2006; van't Wout, Chang and Sanfey, 2010)



as well as by the present study, when participants played in the UG in the MS condition. In fact, it is not excluded that other emotional responses might have entered in this social interaction. It is plausible, for instance, that altruistic feelings and motivations similar to those described by Moll, et al. (2006) with regard to charitable donation, contribute to act in the same way both for oneself and on behalf of another person, by rejecting the unfair behavior. What our findings seem to suggest is that negative emotions are not *always* the key-mechanism underlying the responder's rejections. These emotions might be triggered whenever one's own interest is at stake, and not be the ultimate cause of this behavior.

In the next chapter, I will discuss a study in which the issue is further investigated by using the fMRI, which has helped to disentangle between areas associated with the rejections in the MS and in the TP conditions.

## Chapter 3

# Disentangling self- and fairness- related neural mechanisms: an fMRI study.

### 3.1 Introduction

As discussed in the previous chapter, since the UG is a task which primarily focuses on self-interest, it does not allow disentangling whether rejections are driven primarily by anger or direct personal frustration, elicited by the perception of being treated badly (*Emotional* theory) or by more general considerations about fairness, such as the wish to discourage unfair behavior or violations of social norm. Fairness sensitivity is assumed to emerge from the complex integration of cognitive, emotional and motivational mechanisms (Moll, De Oliveira-Souza and Zahn, 2008) and may lead to behavior which optimizes the aggregate welfare. In this perspective, UG's rejections might be considered a pro-social act, as they lead proposers to less unfair behavior in future interactions (Zamir, 2001; Civali et al., 2010), thereby increasing the overall gain of the population of responders (*Fairness* theory). Crucially, although both *Emotional* and *Fairness* theories consider rejections to be related at least to some extent to one's

emotional state, according to the *Fairness* theory rejections are not considered exclusively emotionally-driven, and might occur even in a neutral – i.e. non-personal – context where anger or personal frustration are not involved. In the previous chapter, it has been described how the dissociation between the behavioral and the emotional responses suggests that rejections might be associated with increased emotional arousal when one's own interest is at stake (see also van't Wout et al., 2006), despite occurring also in a context in which unfairness is directed to a third party (that is, free of or with reduced emotional arousal); therefore, these data suggest that rejections are prevalently *fairness*-driven.

These findings raise the question about the functional properties of brain regions previously associated with the rejection of unfair offers in the UG, such as the anterior insula - AI- and the medial prefrontal cortex -MPFC- (for a review of the neural correlates of the UG, see the Introduction, paragraph 1.2), and whether their neural activity reflects processes related to the self, or processes related to the rejections of an unfair offer *per se* which, consistent with the data presented in chapter 2, should be common to both the myself –MS- and the third party –TP- conditions. The functional Magnetic Resonance Imaging (fMRI) allowed to measuring blood-oxygen-level-dependent (BOLD) signal when healthy participants were engaged in the paradigm described in the previous chapter. For the fMRI analyses, unlike the previous study, a 3 x 2 factorial design was considered, with TASK (UG Rejections, UG Acceptances, Free Win –FW-) and TARGET (Myself –MS-, Third-Party –TP-) as factors, and, consequently, 6 conditions: UR\_MS, rejected trials when playing UG for oneself; UA\_MS, accepted trials when playing UG for oneself; FW\_MS, Free Win task addressing oneself; and, respectively, UR\_TP, UA\_TP, FW\_TP, in which the participants performed the UG and FW tasks on behalf of a third party. Of crucial interest are the functional properties of regions such as the AI and the MPFC. In

particular if these regions code processes related exclusively to fairness, then their activity should be associated with contrasts testing for effects of the UG (as opposed to FW) offers and the rejection (as opposed to acceptances) thereof, which are shared across MS and TP conditions [e.g.,  $(UR\_MS + UA\_MS + UR\_TP + UA\_TP)/2 - (FW\_MS + FW\_TP)$  and  $(UR\_MS + UR\_TP) - (UAm + UA_t)$ ], as both targets share the same amount of unfairness in the UG offers. Alternatively, if neural activity of these regions reflects the direct involvement of the Self in an unfair division, then it should be significantly associated with the interactions of TASK\*TARGET, as reflected in the contrasts testing self-specific increases of neural activity during the assessment of UG offers [i.e.,  $((UR\_MS + UA\_MS)/2 - FW\_MS) - ((UR\_TP + UA\_TP)/2 - FW\_TP)$ ] and the rejection thereof [i.e.,  $(UR\_MS - UA\_MS) - (UR\_TP - UA\_TP)$ ].

## 3.2 Materials and methods

### 3.2.1 Participants

Twenty-three (9 females) subjects took part in the experiment. None of the participants had any history of neurological or psychiatric illness. Written informed consent was obtained from all subjects, who were naive to the purpose of the experiment. The study was approved by the local ethics committee.

### 3.2.2 Task and Stimuli

Task, stimuli and experimental set-up were similar to the ones described in chapter 2. Participants underwent one session of thirty minutes, in which they played as responders in a modified version of the UG. At each trial, participants were told that another participant (i.e., the

proposer), seated at the time of the experiment at a computer station just outside the MRI room, was given a 10 € note, and that he/she had to split this money with the participant (responder). Participants were told that the proposer was free to decide how to divide the money (e.g., he/she could keep most of the money to him/herself; likewise, he/she could divide the money equally, etc.), while knowing that the responder was free to accept or reject the offer and hence decide whether or not the division was going to take place. Subjects performed either the UG or a control task (Free-Win [FW]), in which they accepted/rejected money provided by the computer. In both UG and FW tasks, offers ranged from 1 to 5€-although the instructions were given to induce in the participants the belief that they were interacting with a human fellow as the proposer, participants were presented with offers defined *a priori* by the experimenter-. Furthermore, within UG, we distinguished between trials which were accepted/rejected by the participants (participants seldom reject FW offers -see Results Section and chapter 2-). Finally, in both tasks trials were presented so that offers either addressed participants themselves or a third party. All offers were presented in written (font: Arial, font: 28) and their content varied according to the task employed, the target of the offer and the amount of money offered (e.g., “I offer you 2 euros out of 10”/“I offer to the next participant 2 euros out of 10” [UG]; “You are given 2 € for free”/“The next participant is given 2 € for free” [FW]). All offers were followed by the question “Do you accept?”/“Do you accept on her/his behalf?”

After the whole experimental session an informal debrief was carried out to assess whether participants believed whether offers were genuinely human. None of the participants exhibited doubts about the cover story.

### 3.2.3 Experimental Set-Up

Participants lay supine in the MR scanner with their head fixated by firm foam pads. Stimuli were presented using Presentation 11.0 (Neurobehavioral Systems) and projected to a VisuaStim Goggles system (Resonance Technology). Behavioral responses were recorded by pressing the corresponding keys of an MRI-compatible response device (Lumitouch, Lightwave Medical Industries, CST Coldswitch Technologies).

For each experimental trial, participants were first presented with the offer for 4500 msec, followed by a blank screen ranging from 4750 msec to 6750 msec with an incremental step of 500 msec. The question “Do you accept (on his/her behalf) ?” was then presented for 2000 msec. Trials were followed by an inter-trial interval ranging from 4750 msec to 6750 msec with an incremental step of 500 msec. This trial set-up allowed us to measure putative increases of BOLD signal prior to the delivery of the key-presses, i.e., during the presentation of the offer. This yielded a factorial design, with TASK (UG, FW), TARGET (MS, TP), and GAIN (1, 2, 3, 4, 5 euros) as factors. Each experimental session comprised 105 randomized trials, including 100 experimental trials [2 OFFER x 2 TARGET x 5 GAIN x 5 repetitions] and 5 “null events” in which an empty screen replaced the stimuli.

*fMRI data acquisition.* A Siemens Trio 3-T whole-body scanner was used to acquire both T1-weighted anatomical images and gradient-echo planar T2-weighted MRI images with blood oxygenation level dependent (BOLD) contrast. The scanning sequence was a trajectory-based reconstruction sequence with a repetition time (TR) of 2200 msec, an echo time (TE) of 30 msec, a flip angle of 90 degrees, a slice thickness of 3 mm, and no gap between slices. For each subject, 878 volumes were acquired during the whole experimental session.

#### *3.2.4 Behavioral and imaging data processing*

For each subject, and for each condition, the rejection rate was calculated across all 5 repetitions, and used in a TASK X TARGET X GAIN Repeated Measures ANOVA. Statistical analysis was performed using SPSS 11.5 Software (SPSS Inc., Chertsey UK).

Image processing and statistical analysis were performed using the SPM8 software package (<http://www.fil.ion.ucl.ac.uk/spm/>). For each subject, the first six volumes were discarded. To correct for head motion, the functional images were then realigned to the new first functional image (Ashburner and Friston, 2004), normalized to a template based on 152 brains from the Montreal Neurological Institute (MNI), and then smoothed by convolution with a 8 mm full-width at half-maximum (FWHM) Gaussian kernel.

Data were then fed into a first level analysis using the general linear model framework (Kiebel and Holmes, 2004) implemented in SPM8. On the first level, for each individual subject, we fitted a linear regression model to the data. For the UG only, a distinction between rejected and accepted offers was made (participants seldom reject FW offers – see Results section and Civai et al., 2010). This yielded a 3 (TASK: UG Rejections, UG Acceptances, FW) x 2 (TARGET: MS, TP) factorial design with 6 conditions: UR\_MS, UA\_MS, FW\_MS, UR\_TP, UA\_TP, FW\_TP. For each of these conditions the onset of the offer and the onset of the text string prompting a button press were modeled independently through a stick functions (see Figure 3.1b). For each of the resulting 12 vectors, I also accounted for putative linear changes of neural activity across all repetitions by using the time modulation option implemented in SPM, which creates a new regressor in which the trial order is modulated parametrically. Furthermore, regressors testing the parametric modulation of the factor GAIN were included: distinct regressors were modeled for the two onsets within the trial structure (offer, response), the two levels of TARGET (MS, TP) and for TASK which was UG and FW, but not for different

responses within UG trials. This yielded 32 vectors [12 stick functions + 12 time modulation vectors + 8 gain modulation vectors], each of which was convolved with a canonical haemodynamic response function and associated with a vector describing its first order time derivative. Finally, to account for movement-related variance, six differential realignment parameters were included as regressors. Low-frequency signal drifts were filtered using a cutoff period of 128 sec.

Please notice that, due to intrinsic properties of the bargaining game, regressors testing for specific responses (e.g., UR\_MS) correlate strongly with regressors testing for response-independent effects of offer size (see behavioral results). Following Andrade, Paradis, Rouquette, and Poline (1999), we assume that by inserting two correlated regressors in the same model, the parameters associated with each of them would be estimated on that portion of variance that is not shared with the confound (e.g., effects of rejections/acceptances would be estimated on variance that is independent from the one explained by the size of the offer). Although realizing that modeling both responses and offer size might lead to sensitivity problems, by doing so it is insured that the results (if any) could be uniquely interpreted, thus ruling out potential confounding effects of the correlated regressor (Andrade et al., 1999).

The first level analysis of each subject yielded images describing the parameter estimates associated with each of the vectors modeled. Of key interest for the current purposes are those parameter estimates associated with the 6 conditions of our 3 x 2 design, exclusively when the offer was presented (but not when the response was triggered). These images were then fed into a second-level flexible factorial design with a within-subject factor of six levels using a random effects analysis (Penny and Holmes, 2004). The effects of the offer size were also assessed by feeding, in a similar second-level flexible factorial design, the four parameter estimates testing



for the parametrical modulation of the factor GAIN (also in this case, the parameters were only those associated with the onset of the offer, and not with the onset of the string prompting the button presses).

For each activated region, the percentage signal changes were calculated over the local maxima parameter estimates using the MarsBar toolbox (Brett et al., 2002). Further statistical analyses were performed over the extracted percentage signal change values to further investigate the functional properties of the areas of activation. This statistical analysis was performed with SPSS 11.5.

### **3.3 Results**

#### *3.3.1 Behavioral results*

One subject never rejected UG offers, neither in the MS nor in the TP condition, whereas another subject never rejected third-party UG offers only. The remaining 21 subjects rejected UG offers in both MS and TP conditions. For each of the 23 subjects and for each condition, the rejection rates were calculated across all 5 repetitions, and used in a 2 TASK (UG, FW) x 2 TARGET (MS, TP) x 5 GAIN (1-5 €) Repeated Measures ANOVA. Results indicated a significant main effect of TASK ( $F(1, 22) = 123.89, p < 0.001$ ), with the UG leading to a larger number of rejections than the FW, as well as a main effect of GAIN ( $F(4,88) = 58.73, p < 0.001$ ), with lower offers being rejected more often than higher offers. These effects were, however, driven by a TASK \* GAIN interaction, which was also significant ( $F(4,88) = 63.44, p < 0.001$ ), suggesting that lower offers were rejected significantly more often than higher offers in the UG but not in the FW (see Figure 3.1). None of the remaining effects of the ANOVA was significant.

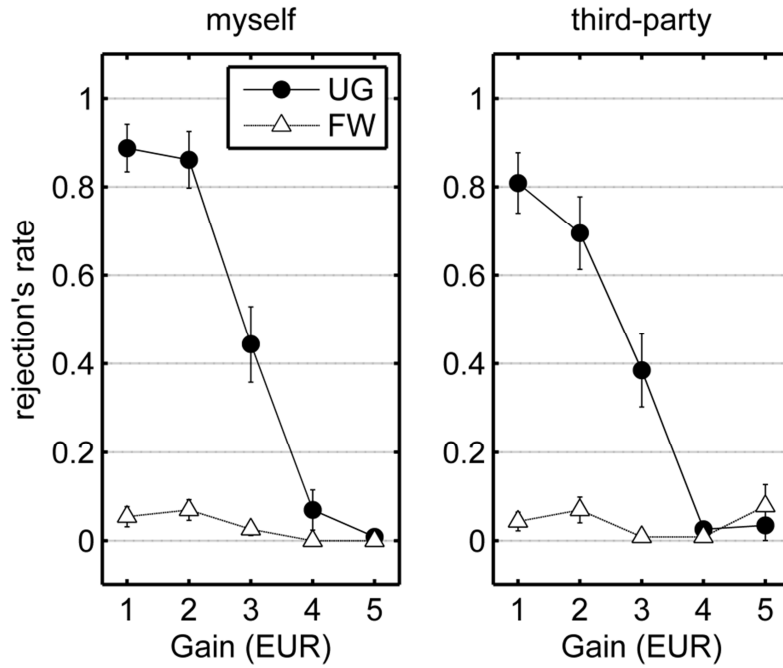


Figure 3.1. Behavioral results. Rejection Rates are plotted as a function of Gain in both UG (black circles) and FW (white triangles) tasks and MS and TP conditions.

### 3.3.2 Neural Activations

Unless stated otherwise, we report exclusively areas of activation which survived an extent threshold  $> 176$  consecutively activated voxels (corresponding to a  $p < 0.05$ , corrected for multiple comparisons across the whole brain), with an underlying height threshold of  $t > 3.17$  (corresponding to  $p < 0.001$ , uncorrected). Please see Table 1 for the full set of results.

Table 3.1. Voxels showing significant increases of neural activity associated with TASK, TARGET and TASK\*TARGET interaction. All clusters survived a threshold corresponding to  $p < 0.05$ , corrected for multiple comparisons across the whole brain, with an underlying height threshold corresponding to  $p <$

0.001 (uncorrected). Only contrasts yielding significant activations are reported. Coordinates are in standard MNI space.

REGION	SIDE	Coordinates		
		X	Y	Z
1: Main effect of TASK: UG > FW				
$(URm + UAm + URt + UAt)/2 - (FWm + FWt)$				
Supramarginal Gyrus	R	42	-34	38
Precuneus		10	-64	38
Calcarine Gyrus		10	-62	10
Supramarginal Gyrus	L	-40	-36	38
Precuneus		-8	-62	36
Calcarine Gyrus		-10	-66	10
Midbrain/PAG	R	10	-4	-10
	L	-6	-12	-10
Anterior Insula	L	-34	16	0
Supplementary Motor Area	L	-2	18	44
Anterior Cingulate Cortex		-8	32	14
Precentral Gyrus	L	-40	-6	52
Anterior Insula	R	30	22	2
Middle Frontal Gyrus	R	36	10	54
Middle Temporal Gyrus	L	50	-62	-12
Inferior Occipital Gyrus	L	36	-84	-12
	R	-30	-78	-8
Inferior Frontal Gyrus	R	48	4	26

<b>2: Main effect of TASK: Rejected &gt; Accepted Ultimatum Game offers</b>				
$(UR\_MS + UR\_TP) - (UA\_MS + UA\_TP)$				
Midbrain/PAG	L	-6	-26	-6
		-2	2	-10
<b>3: Main effect of TARGET: MS&gt;TP</b>				
$(UR\_MS + UA\_MS + FW\_MS) - (UR\_TP + UA\_TP + FW\_TP)$				
Middle Orbital Gyrus		-2	38	-6
Superior Medial Gyrus	R	6	54	2
Inferior Frontal Gyrus	R	48	46	-6
		26	22	-18
<b>4: Main effect of TARGET: TP&gt;MS</b>				
$(UR\_TP + UA\_TP + FW\_TP) - (UR\_MS + UA\_MS + FW\_MS)$				
Lingual Gyrus		14	-78	-6
Superior Occipital Gyrus	R	26	-82	20
Lingual Gyrus	L	-12	-82	-14
Superior Occipital Gyrus		-16	-92	24
Inferior Parietal Cortex	L	-44	-68	28
Middle Temporal Gyrus	L	-60	-8	-14
<b>5: OFFER*TARGET interaction: UG &gt; FW, specifically for MS</b>				
$((UR\_MS + UA\_MS)/2 - FW\_MS) - ((UR\_TP + UA\_TP)/2 - FW\_TP)$				
Superior Medial Gyrus	R/L	0	58	8

*Main effects.* I first tested for regions exhibiting significant increases of neural activity for UG as opposed to FW [i.e.,  $(UR\_MS + UA\_MS + UR\_TP + UA\_TP)/2 - (FW\_MS + FW\_TP)$ ].

Bilateral activations were found at the level of the calcarine gyrus, the inferior occipital gyrus, the inferior parietal cortex, and the anterior aspect of the insula. Activation of the midbrain, of the anterior cingulate cortex, of the left precentral gyrus, the right middle frontal and temporal gyri was also found (Figure 3.2).

I next tested for increases of neural activity associated with offers addressing oneself (irrespective of whether these were UG or FW) as opposed to offers addressing a third-party [i.e.,  $(UR\_MS + UA\_MS + FW\_MS) - (UR\_TP + UA\_TP + FW\_TP)$ ]. Such increases were found at the level of the medial prefrontal cortex (Figure 3.4b, violet cluster), including the most ventral part (14 mm below the inter-commissural line), and extending to the superior medial gyrus (2 mm above the inter-commissural line). Further activation was found at the level of the inferior frontal gyrus.

Moreover, regions exhibiting increased neural activity when about to reject (as opposed to accept) UG offers were tested [i.e.,  $(UR\_MS + UR\_TP) - (UA\_MS + UA\_TP)$ ]. This contrast revealed activation of midbrain regions, over and around the midbrain cluster revealed by the previous analysis. Following Sanfey et al. (2003), who reported the AI as most strongly active when participants were about to reject (rather than accept) UG offers, we also expected in our study the AI to be significantly associated with UG rejections. Therefore, a restricted analysis on the bilateral insula (AAL Atlas – Tzourio-Mazoyer et al., 2002) found a significant activation in the inferior aspect of the left AI ( $x = -36, y = 16, z = -4, t(107) = 4.05, p < 0.05$ , family-wise corrected for the region of interest). Figure 3.3a (yellow cluster) displays this increase in neural activity, which is  $< 5$  mm distant from the left anterior insular region (MNI-converted coordinates  $x = -33, y = 14, z = 0$ ) reported by Sanfey et al. (2003). Please notice that, in this analysis both participants' responses and response-independent linear effects of the factor GAIN

in the UG were modeled (see Methods section); thus the parameters associated with rejections in the insula and midbrain should describe specific response effects, over and above those associated with the offer size (Andrade et al., 1999). In depth analysis was carried out on the percentage signal changes extracted from the cluster local maxima (see Figure 3.3c, left graph) to assess putative TARGET-modulations in this region. As not all participants rejected UG offers in both MS and TP condition (see behavioral results section), we used a Linear Mixed Model (McCulloch et al., 2008) which is more robust than a traditional ANOVA against missing cells. The model included RESPONSE (Accept, Reject) and TARGET (MS, TP) as fixed factors, and Subjects as random factor. A compound symmetry covariance structure was specified. We found a significant main effect of RESPONSE ( $F(1, 64.69) = 14.97, p < 0.001$ ), reflecting an overall increase of insular activity when UG offers are about to be rejected, but neither a main effect of TARGET ( $F(1, 63.62) = 1.02, n.s.$ ) nor a RESPONSE \* TARGET interaction ( $F(1, 63.62) = 0.19, n.s.$ ) was observed. Finally, paired-samples  $t$ -tests confirmed that the insular RESPONSE effect survived analysis when considering separately UG offers addressing oneself (URm – UAm:  $t(21) = 3.28, p < 0.01$ ) and the third-party (URt – UAt:  $t(20) = 2.15, p < 0.05$ ).

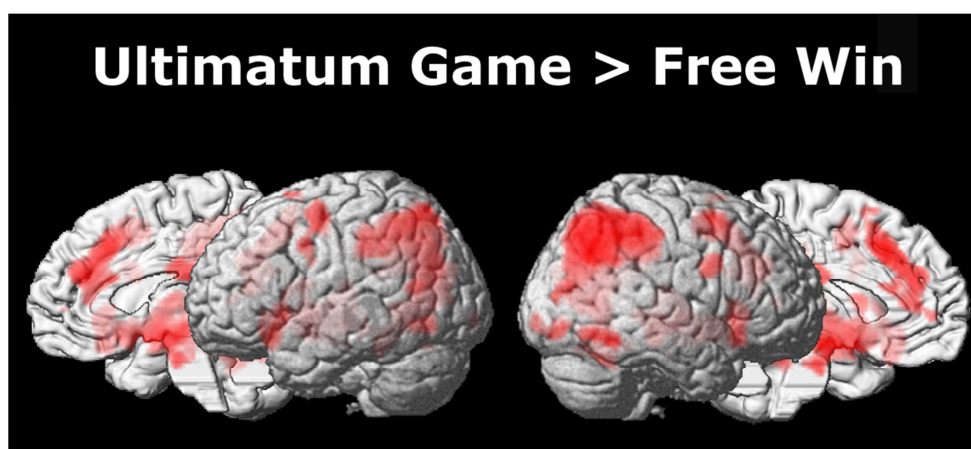


Figure 3.2. Surface renderings of the functional contrasts testing regions exhibiting a larger neural activity when subjects were engaged in UG, rather than FW.

*Interactions.* I tested for significant increases of neural activity associated with the UG (but not the FW), exclusively when offers addressed oneself and not the third-party [i.e.,  $((UR_m + UAm)/2 - FW_m) - ((UR_t + UAt)/2 - FW_t)$ ]. This analysis isolated the medial prefrontal cortex. Figure 3.3b displays this region (green cluster) together with the adjacent (and partially-overlapping) region revealed by the analysis of the main effect of TARGET (violet cluster), thus showing that the region revealed by the interaction term lies more dorsal and frontal with respect to the region involved in processing offers addressing oneself. In depth analysis was carried out on the percentage signal changes extracted from the cluster local maxima in order to assess RESPONSE effects in this region. Therefore, a Linear Mixed Model was carried out with RESPONSE (Accept, Reject) and TARGET (MS, TP) as fixed factors and Subjects as random factor. A compound symmetry covariance structure was specified. No effect of RESPONSE ( $F(1, 64.44) = 0.06, n.s.$ ) was found, but a significant main effect of TARGET ( $F(1, 63.60) = 14.81, p < 0.001$ ), reflecting an overall increase of MPFC activity when UG offers addressed oneself (as opposed to a third party), and a significant RESPONSE \* TARGET interaction ( $F(1, 63.60) = 5.88, p < 0.05$ ), revealing that the self-specific increase of MPFC activity in the UG was larger during rejections than acceptances (see Figure 3.3c, middle graph).

I also tested for regions exhibiting specific increases for rejected (as opposed to accepted) UG offers, specifically when they addressed oneself, rather than the third party:  $[(UR_m - UR_t) - (UAm - UAt)]$ . However, no significant effect was found, neither when testing for the whole brain, nor when restricting our interest on MPFC and the left insula. We tested also for significant increases of neural activity for UG offers (and the rejection thereof) specifically when

these address the third-party [i.e.,  $((UR_t + UA_t)/2 - FW_t) - ((UR_m + UA_m)/2 - FW_m)$  and  $(UR_t - UR_m) - (UA_t - UA_m)$ ]. No significant effects were found.

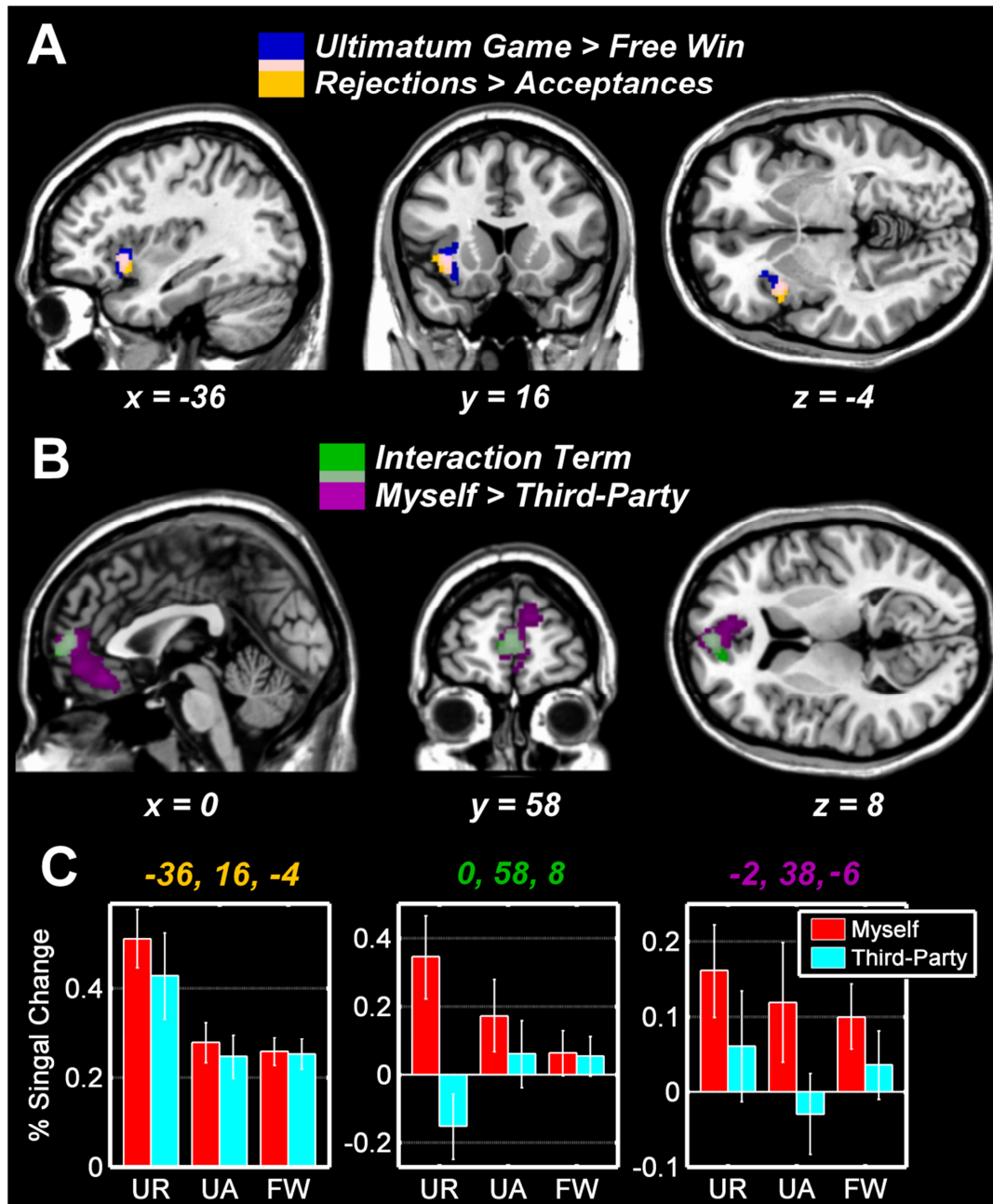


Figure 3.3. (A) Sagittal ( $x = -36$ ), coronal ( $y = 16$ ) and axial ( $z = -4$ ) sections displaying the functional contrast testing  $UG > FW$  (blue activation) and, within  $UG$  offers, the contrast testing  $Rejections >$



Acceptances (yellow activation). Light pink activations refer to regions significantly associated with both contrasts. **(B)** Sagittal ( $x = 0$ ), coronal ( $y = 58$ ) and axial ( $z = 8$ ) sections displaying the functional contrast testing the interaction term (green activations) and  $MS > TP$  (violet activation). Light green activations refer to regions significantly associated with both contrasts. **(C)** The parameter estimates associated with representative voxels of the activated areas are displayed together with S.E.M. error bars. Red bars refer to offers addressing oneself, whereas cyan bars refer to offers addressing a third-party.

*Simple effects.* We focused our attention only on TP. We first investigated for effects of the Ultimatum Game  $[(UR\_TP + UA\_TP)/2 - FW\_TP]$ , and isolated, reminiscently to the activation displayed in Figure 3.2, the intraparietal sulcus (extending to the inferior parietal cortex) bilaterally and midbrain structures over and around the substantia nigra and the red nucleus. Interestingly, the midbrain activation overlaps not only previous results associated with the main effect Ultimatum Game  $>$  Free Win, but also the midbrain region associated with Rejections  $>$  Acceptances. We then tested for increases of neural activity associated to rejections, rather than acceptances ( $UR_t - UA_t$ ). No suprathreshold activation was found

*Parametrical modulation of GAIN.* No region showed significant effects associated with the parameters testing the parametrical modulation of the factor GAIN, at least when using rigorous correction for multiple comparisons for the whole brain. We then restricted our hypothesis on the ventral portion of the MPFC isolated when testing the main effect of TARGET (see Figure 3b, violet blob). We therefore extracted the four parameters testing the GAIN effect from the cluster's representative voxel (-2, -38, -6) and subjected them to one-sample t-tests in order to assess significant deviations from 0. Only the parameters testing effects of offers size in the  $FW\_MS$  was found to be significantly larger than 0 ( $t(22) = 3.10$ ,  $p < 0.01$ ). This was not the case for the other three parameters ( $UG\_MS$ :  $t(22) = 0.99$ ;  $UG\_TP$ :  $t(22) = 0.45$ ;  $FW\_TP$ :  $t(22) =$

0.98). The same analysis was carried out on the peak from the inferior frontal activation (48, 46, -6) isolated as well through the main effect of TARGET (see Table 1). In this case, however, none of the parameters were significantly different from 0 (UG\_MS:  $t(22) = 0.07$ ; UG\_TP:  $t(22) = -0.69$ ; FW\_MS:  $t(22) = -0.38$ ; FW\_TP:  $t(22) = 0.59$ ).

### 3.4 Discussion

I employed the modified Ultimatum Game (UG) paradigm described in chapter 2, in which participants played either for themselves (MS) or on behalf of a third-party. Using fMRI, I found the anterior insula involved in dealing with unfair offers affecting both oneself and others, as revealed by the contrast testing for increased neural activity associated with rejections (as opposed to acceptances) in both the MS and TP condition. Instead, the middle-anterior portion of the MPFC was recruited exclusively when the unfair offers were related to oneself only. These data converge with, but also extend, previous studies: I have not only mapped the neural mechanisms underlying people's reaction to unfairness, but we also disentangled those processes reflecting judgments related to unfair behavior (fairness), from those related to the cognitive/emotional processes when oneself needs to deal with unfair behavior (self-effect).

#### 3.4.1 *Self-Specific Neural Networks*

Studies in the field of economics implicated both middle-anterior and ventral portions of the MPFC in tasks in which participants assessed the value of potential outcomes (see, Amodio and Frith, 2006, as review): for example, Knutson, Taylor, Kaufman, Peterson, and Glover (2005) associated the activity of part of the MPFC ( $z \approx -6$ ) with the computation of expected monetary value, whereas Coricelli and colleagues testing both healthy subjects with fMRI (Coricelli et al., 2005) and brain damaged patients (Camille, et al., 2004; Larquet, Coricelli,

Opolczynski, and Thibaut, 2010) implicated a more ventral part of MPFC (extending to  $z \approx -14$ ) in anticipated regret associated with monetary decision. Neuropsychological studies testing the classical UG task report results in line with those from neuroimaging studies: whereas Koenigs and Tranel (2007) described patients with MPFC damage (from ventral to middle-anterior portions) more prone to reject unfair offers, Moretti et al. (2009) limit Koenigs' findings to the case in which bargaining offers are described as abstract sums to be received later, rather than visible and immediately-available banknotes, thus suggesting a MPFC role in representing the value of abstract outcomes (and, therefore, of the economic benefit of unfair bargaining). The role of MPFC in economic choice is, however, not limited to assessing the value of possible outcomes, rather MPFC is also suggested to be involved in decision making and mentalizing. For instance, Coricelli and Nagel (2009) associated activity of parts of the MPFC ( $z \approx -9, +24$ ) with high-level reasoning as assessed with the *Beauty Contest*, a game which assesses the subject's ability to develop strategies based on representations of competitors' putative choices. Crucially, in all these studies the role of MPFC is established by tasks exposing directly participants to potential gains or losses. I tested MPFC sensitivity to potential monetary gain in a TP condition, and found activation of the middle-anterior and ventral MPFC (extending ventrally to  $z \approx -14$ ) whenever participants themselves (but not a Third-Party) received money (irrespective of whether the received amount of money resulted from a social interaction (UG) or was given "for free" (FW)). Consistently, the activity of the ventral MPFC increased linearly with the amount of money participants (but not the third-party) gained in the FW task, thus insuring the involvement of this region in personal gain rather than in mere self-reflective processing. Please notice that no linear effect of offer size was observed in this region during UG offers addressing oneself.

Indeed, at variance with FW, in UG differences in offer size do not reflect exclusively the amount of money gained, but rather the amount of unfairness in the proposer's choices.

The functional properties of the middle-anterior aspect of the MPFC are heterogeneous and involve the co-occurrence of cognitive, emotional and social processes. First of all, the middle MPFC responds to emotional events: for example, extensive MPFC activation was found during passive processing of emotional pictures or (Lee, et al., 2004) or words (Beauregard, et al., 1997), or in studies in which participants were asked to rate explicitly the emotional content of pictures (Lane, Fink, Chau, and Dolan, 1997; Gusnard, Akbudak, Shulman, and Raichle, 2001; Dolcos, LaBar, and Cabeza, 2004). Furthermore, Ochsner, et al. (2004) reported activation of the middle-anterior MPFC ( $z \approx 8$ ) when participants judged their own affective state, whereas Peelen, Atkinson, and Vuilleumier (2010) reported a slightly superior region ( $z \approx 21$ ) coding emotional states irrespective of whether these are experienced through voice, facial expression or body language. The middle-anterior MPFC ( $z$  from 2 to 17) has also been implicated in self-reflection, as suggested by studies engaging participants in self-judgments about traits/adjectives or episodic memory (Zysset, Huber, Ferstl, von Cramon, 2002; Kelley, et al., 2002) and in mentalizing ( $z$  from 3 to 48), that is when participants were asked (through storyboard or pictorial tasks) to assess others' mental states (Goel, Grafman, Sadato, and Hallett, 1995; Saxe and Powell, 2006). Finally, middle-anterior MPFC activation ( $z$  from -12 to 24) has been reported for moral judgments and reasoning (Greene, Sommerville, Nystrom, Darley, and Cohen, 2001; Moll et al., 2002; Heekeren, Wartenburger, Schmidt, Schwintowski, and Villringer, 2003; Green, Nystrom, Engell, Darley, and Cohen, 2004; Schaich Borg, Hynes, Van Horn, Grafton, and Sinnott-Armstrong, 2006), whereas its dysfunction has been described leading to impaired moral behavior (Koenigs, et al., 2007; Ciaramelli, Muccioli, Ladavas, and di Pellegrino, 2007;

Krajbich, Adolphs, Tranel, Denburg, and Camerer, 2009; Moll, et al., 2011). Overlaps between these different cognitive processes have often been suggested: indeed, whereas Jenkins, Macrae, and Mitchell (2008) describe mentalizing effects as reflective of self-referential processing, other studies suggest a privileged MPFC role in mentalizing about others' emotional states (Hynes, Baird, and Grafton, 2006; Shamay-Tsoory, Tibi-Elhanany, and Aharon-Peretz, 2006; Gilbert, et al., 2006). Likewise, self-referential and emotional processing in MPFC are often confounded one another, as performance in emotional rating might be a self-referential task (Amodio and Frith, 2006) or the Self can be considered as an emotional entity *per se* (Modinos, Ormel, and Aleman, 2009).

I here report an activation of the middle-anterior portion ( $z \approx 8$ ) of the MPFC (Figure 4b, green cluster) which may reflect some of these complex cognitive processes. Indeed, as for the case of the more ventral portion of the MPFC, it was modulated specifically in the MS (but not TP) condition. At variance with the ventral MPFC, the modulation was restricted to the UG (but not the FW), that is only when a potential monetary gain was the result of a social interaction. Furthermore, analysis on the extracted percent signal changes revealed a larger MS > TP effect in those UG trials that were about to be rejected than accepted, suggesting an additional recruitment of this area when facing self-directed unfair behavior. This result is reminiscent of data described in the last chapter, which showed enhanced skin conductance responses associated with rejection (rather than acceptance) of unfair UG offers in the MS condition. Taken together the data suggest that this middle-anterior MPFC activity might be related to emotional arousal evoked by an unfair offer related to the self. Amodio and Frith (2006) suggested that value-related representations in the ventral MPFC extend the more anterior (and superior), the more complex they become, and that they integrate with socio-affective processes. In this study, this

progression is showed in the context of monetary gain/loss when participants are personally involved in the bargaining.

#### *3.4.2 Fairness-Related Neural Networks*

A significant increases of neural activity during the UG (with respect to the FW) in the right anterior cingulate cortex, the right middle frontal gyrus, the precuneus bilaterally and the inferior parietal lobe bilaterally, extending to the intraparietal sulcus, was found. In line with Sanfey et al (2003), these activations are likely to reflect a large number of processes underlying task performance, amongst which the increased calculation effort (Le Clec', 2000), the cognitive conflict (Botvinick, Nystrom, Fissell, Carter, and Cohen, 1999; MacLeod and MacDonald, 2000), and the need of maintaining the task's goal in working memory (Miller and Cohen, 2001) all can be subsumed.

The present study implicates the left anterior insula not only in the analysis testing effects of UG (as opposed of FW) in both MS and TP condition, but also in the analysis testing rejections (as opposed to acceptances) of UG offers. In both analyses, the voxels isolated were close (i.e., within < 5 mm) to the ones previously reported by Sanfey et al. (2003) as sensitive to UG rejections (corresponding to our MS condition), or by Tabibnia et al. (2008) as more recruited in those participants who often rejected unfair offers. Furthermore, activations close to this region (< 8 mm) were involved also in anticipation of one's monetary gain/loss (Ernst, et al. 2004; Knutson, Taylor, Kaufman, Peterson, and Glover, 2005; Preuschoff, Bossaerts, and Quartz, 2006; Knutson, Bhanji, Cooney, Atlas, and Gotlib, 2008; Engelmann, Capra, Noussair, and Berns, 2009). Given that previous studies reported this portion of the anterior insula (< 8 mm) involved in negative experiences, such as disgust (Shapira et al., 2003), pain (Peyron et al., 1999;

Hui et al., 2000; Mohr, Leyendecker, and Helmchen, 2008) or thirst (Denton et al., 1999; de Araujo, Kringelbach, Rolls, and McGlone, 2003), it might be argued that many economical tasks are emotionally-grounded, and that the activation of the anterior insula in these tasks is a physiological marker of negative emotional involvement. Thus, these results might be considered, contrary to the evidences reported in chapter 2, consistent with TP rejections as negative emotional as rejections associated with the MS condition (see also Sanfey et al., 2003; Chang and Sanfey, 2009). This view, however, does not acknowledge that the same portion of the anterior insula ( $< 8$  mm) has also been associated with processing positive events (Malhi, et al. 2007), or cognitive processes which are not necessarily emotionally-grounded, such as motor control (Aramaki, Honda, Okada, and Sadato, 2006) or attention allocation (Steel et al., 2001; Kelly, et al., 2004; Durston, Mulder, Casey, Ziermans, and van Engeland, 2006; Chikazoe et al., 2009). Furthermore, previous studies suggest that the anterior insula activity associated with negative emotions is more than a physiological response of one's emotional/somatic state, but rather reflects explicit awareness of what one's and others' states might be (Craig, 2003, 2009; Singer, Critchley, and Preuschoff, 2009; Lamm and Singer, 2010): thus, not only a physiological index of emotional arousal – such as electrodermal activity and heart beat – but a structure involved in explicitly monitoring these indexes (Critchley, Wiens, Rotshtein, Ohman, and Dolan, 2004); not only a region involved in detecting a noxious event, but also in coding its perceived unpleasantness (Rainville, Duncan, Price, Carrier, and Bushnell, 1997; Craig, Chen, Bandy, and Reiman, 2000), or its predictability from preceding cues (Porro et al., 2002; Atlas, Bolger, Lindquist, and Wager, 2010); not only responsive to negative visual stimuli, but involved in explicitly assess their negative (e.g., painful) content (Lamm, Nusbaum, Meltzoff, and Decety, 2007; Gu and Han, 2007).

Recent accounts suggest that the anterior insula integrates information about modality-specific feelings with cognitive processes, individual preferences and contextual information in order to promote behavioral responses (Singer et al., 2009; Lamm and Singer, 2010). In this perspective this region is an ideal candidate for mediating *fairness*-related behavior which emerges from the integration of cognitive, emotional and motivational mechanisms (Moll et al., 2008). Indeed, anterior insula activity to stimuli depicting people in pain has been shown to be affected by contextual information about these people, such as their status in the community (Decety, Echols, and Correll, 2009) or their fairness in economic scenarios (Singer et al., 2006). Furthermore, this region mediates punishments of unfair behavior in social interactions not only in the UG (Sanfey et al., 2003; Tabibnia et al., 2008): for instance, Rilling et al. (2008) implicated coordinates proximal to ours ( $< 5$  mm) in unreciprocated (as opposed to reciprocated) cooperation during the *Prisoner's Dilemma*, King-Casas et al. (2008) associated the anterior insula ( $< 10$  mm) with borderlines patients' inability to maintain cooperation in a *Trust Game*, whereas Strobel et al. (2011) reported activations the same region ( $< 5$  mm) when participants sanctioned unfair offers in the *Dictator Game*. In almost all these studies, the economical games affected directly participants' gain/loss, thus leaving open the possibility that the insular activity they reported was reflective of emotional reactions to unfair treatment or concerns about one's welfare. Strobel et al. (2011) found insular activity even when participants punished TP unfair offers in the Dictator Game, thus implicating this region also when the unfairness sanctioned was directed to someone else. However, TP punishments in Strobel et al. (2011) were costly for participants, thus still leaving open the possibility that the insula responded to emotional concerns about one's money loss. This is not the case of this study in which participants choices in the TP did not affect their own pocket. Thus, it provides evidence in favor of the anterior



insula mediating *fairness*-related behavior. In this perspective, previous studies already described the anterior insula as endowed with “shared” properties, that is responding both when an emotional event is felt directly and perceived in others. This was the case of the experience of disgusting or pleasing tastes/odors (Wicker et al., 2003; Jabbi, Swart, and Keysers, 2007) or pain (Lamm, Decety, and Singer, 2011). However, to the best of our knowledge no study had thus far described shared properties in the insula when facing complex emotional/motivational responses, such as those involved in promoting pro-social behaviors, not only when this benefits oneself, but also when it benefits someone whom participants know nothing about.

It still remains to be elucidated whether the *fairness*-related behavior associated with the anterior insula reflects a moral act, motivated by the wish of sanctioning an intentional unfair action, or inequity aversion, motivated by the wish of preventing an unfair division from taking place, irrespectively of its moral salience (Fehr and Schmidt, 1999). Both factors seem to contribute to rejections in the MS condition, as offers are rejected also when unfair divisions are randomly generated by a computer, although not as frequently as in the case of human-generated divisions (e.g., Sanfey et al., 2003; van’t Wout et al., 2006). Reminiscently to the case of randomly-generated offers, rejections of TP offers can hardly be interpreted as punishments, as the person affected by participants’ choice is not the one responsible of the unfair division and, therefore, is morally unaccountable. In this perspective, an interpretation of insular activity in terms of inequity aversion seems, at the present state, the most plausible. It should be mentioned, however, that an interpretation in terms of moral punishment is still possible, but only if we assume that the target of the punishing act is not a specific unfair individual, but the overall population of proposers, including those who might behave irrationally in subsequent

experimental sessions. Future studies will be conducted to investigate more thoroughly the role played by the insula in moral considerations and inequity aversion.

## Chapter 4

# Driving principles in decision-making: the role of abstract social rules.

### 4.1 Introduction

Despite the fact that, in its classical formulation, Game Theory (GT) fails to predict behavior in the UG, its predictions provide a useful benchmark. The classical formulation of GT (von Neumann Morgenstern, 1944) tried to predict the behavior of rational players, each choosing an action in a game where the profile of chosen actions delivers a consequence. Consequences have utility for players: in the classical concept of Nash, players try to maximize this utility taken the behavior of the others as given. To better understand what the GT predicts in the UG, let us take for a moment the amounts in the description of the extensive form game as measuring utility, rather than money. If we accept this identification, then GT predicts very little if we consider the equilibrium concept to be Nash equilibrium. For example, take an arbitrary amount  $x$ . A strategy profile (that is, a plan assigned to every player, describing a choice in every possible contingency) where the responder accepts an offer if and only if it is larger or equal to  $x$ ,

and the proposer offers  $x$ , is a Nash equilibrium. If one adopts a more restrictive concept (e.g., the Subgame perfection, Selten, 1975)<sup>2</sup>, then equilibrium of this game of perfect information is found by backward induction and is unique if a zero offer is not allowed. In this equilibrium, the responder accepts any offer and the proposer offers the minimum amount. A large number of experiments, beginning with Güth et al. (1982), used money payments and found behavior substantially different from the offer of a minimal amount by the proposer and acceptance of any positive amount by the responder. (If or As) Money is not utility, thus the comparison with the behavior predicted by SP equilibrium has to be interpreted. Is there a discrepancy between game theoretic predictions and behavior of subjects? There are three possible ways to answer this question.

1. A first answer takes monetary amounts as utilities, and explains the behavior of the proposer offering an amount larger than the minimum as due to failure to expect full rational behavior by the responder, who might reject a low offer (Weg & Zwick, 1994). This explanation is consistent with Nash equilibrium prediction, not subgame perfection. In particular it fails to explain the behavior of responders who reject low offers and receive a zero amount, when in the

---

<sup>2</sup> Here is an example, that can be found in The Concise Encyclopedia of Economics, in order to clarify the difference between Nash equilibrium and Selten's Subgame perfection: the Chain Store Paradox. Imagine that firm A has a number of chain stores in various locations and that firm B contemplates entering in one or more of these locations. If firm A threatens a price war, then firm B might be dissuaded from entering, not just in a particular market, but in any of A's markets. In that case, it could well be worthwhile for A to threaten and, indeed, to carry out a price war in a single market. Knowing this, B does not enter. This is a Nash equilibrium. But Selten also observed that another Nash equilibrium was for B to enter. Why? B would realize that A would have losses in each market in which B entered and A carried out a price war. These losses, cumulatively, would not be worthwhile. By looking forward and reasoning backward (backward induction), B realizes that A will not carry on a price war, and therefore B enters. Which Nash equilibrium is the "right" one? Selten argued that it is the one where B enters because B thought through the whole sequence and realized that, from A's viewpoint, a price war would be irrational. B's strategy of entry and A's strategy of avoiding a price war are "subgame perfect."

same situation acceptance would give them a positive amount.

2. A second way to explain observed behavior of experimental subjects and rescue game theoretic predictions is to postulate that monetary payments to a subject are very different from his utility for that outcome. This seems a natural route. Consider, for instance, the classical game of chicken: two players drive their car one against the other, and they can choose between swerve (S) and go straight (G). The utility that they derive does not of course depend only on the trajectory of their car: if a player chooses S, his utility is very different when he shares with the other the embarrassment for choosing S, and when he suffers the unique shame of being the only one to choose S. A full game theoretic approach to the problem of predicting behavior would require that preferences over outcomes of all players are elicited independently of the game. This can be done, for example, by asking the players to choose among random devices assigning payments to all players. Once this is done (for example, once we have measured how much in the chicken game the subject prefers the outcome where both choose S to the one where he chooses S and the other G), then we can test the equilibrium prediction. In the case of the UG, one might conjecture that players may care about the payment to others (e.g., Guth 1988). Specifically, theories of inequality aversion (Bolton, 1991; Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000) assume that utility of players are equal to the monetary amount received, minus a term which is proportional to the difference between the amounts of the two players. This term which is subtracted may be of different sizes depending on whether the monetary amount received by the subject is smaller or larger than the other's.<sup>3</sup> These choices are then used as

---

<sup>3</sup> From Fehr & Schmidt (1999), p.882: "Formally, consider a set of  $n$  players indexed by  $i \in \{1, \dots, n\}$ , and let  $x = x_1, \dots, x_n$  denote the vector of monetary payoffs. The utility function of player  $i \in \{1, \dots, n\}$  is given by  
 (1)  $U_i(x) = x_i - \alpha_i \frac{1}{n-1} \sum_{j \neq i} \max\{x_j - x_i, 0\} - \beta_i \frac{1}{n-1} \sum_{j \neq i} \max\{x_i - x_j, 0\}$ ,  
 where we assume that  $\beta_i \leq \alpha_i$  and  $0 \leq \beta_i < 1$ . In the two-player case (1) simplifies to

preferences to predict behavior in the UG. However the testing of the theory does not proceed by first establishing preferences over outcomes (of the player and the others), for example by offering choices of lotteries, and then using these elicited preferences to predict behavior in the UG. Instead, the behavior in the UG is used for both deriving preferences *and* testing behavior. Without a restriction given by an independent elicitation of preferences, rejecting of theories of inequality aversion becomes very hard to falsify. More precisely, if one assumes inequality averse preferences, then the only restriction on behavior in the UG is that rejection as a function of the share of the responder is weakly decreasing until the 50-50 split is reached, and weakly increasing for higher values. One important prediction that is common to all specifications of inequality averse preferences is that these are preference over outcomes, and hence are independent of what caused the outcomes. For example, if inequality aversion is what drives the choice of players in the UG, then the responder should always reject a low offer, even though the offer is decided by chance.

3. A third approach, which I have widely discussed in the previous chapters, is to dispense with GT, uses a psychological, rather than rational choice, perspective and invokes negative emotions, such as frustration, as the ultimate cause of rejections.

In this chapter I will provide results from experiments in which I used variations of the classical UG that are not consistent with any of the theories reviewed here, and suggest instead a different interpretation of behavior of subjects in these experimental games. I propose the following hypotheses:

---


$$(2) U_i(x) = x_i - \alpha_i \max\{x_j - x_i, 0\} - \beta_i \max\{x_i - x_j, 0\}$$

The second term in (1) or (2) measures the utility loss from disadvantageous inequality, while the third term measures the loss from advantageous inequality.”

1. Behavior in the UG is driven by cognitive factors that implement an abstract moral rule of equal splitting. When asked to split an endowment, or to accept or reject a proposed allocation, subjects are cued to use a moral rule about allocations. When there is no sufficient reason to deviate from equal split (because, for instance, there is no merit in being a proposer), then the default rule of equal splitting applies.

2. This rule, as a moral rule, applies when intentions, not outcomes, are relevant.

3. This rule applies in all situations and roles, because it is a moral rule rather than a preference. Thus, it can be extended to predict the behavior of players for whom the moral rule is relevant (for example, third parties called to decide for others).

#### *4.1.1 Experimental method*

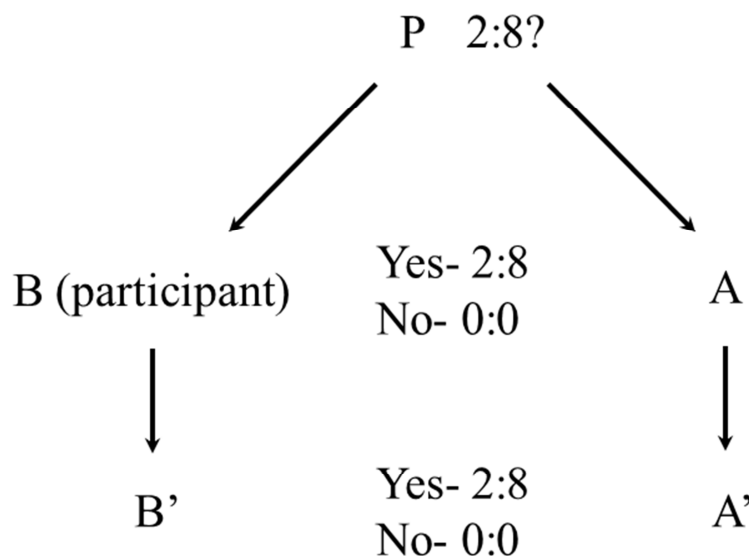
The experiments have been carried out in a completely controlled environment, in which participants believed they were playing together with other participants who had different roles (*proposers*, third parties), when they were actually dealing with offers decided a priori by the experimenter, in the same fashion as in other psychological studies (Civai et al., 2010; Sanfey et al., 2003; van't Wout et al., 2006; Crockett et al., 2008). In these studies the participant knew that the final payoff reflected the proportion of money she had decided to accept. On this point, the instructions stated:

*[...] For instance, if P offers for ten times 1 euro to you and 9 euros to A, and you always accept, you end up with 10 euros and A ends up with 90 euros, which, in percentage, will be 1 euro and 9 euros.*

Participants were eventually paid a fix amount of money, i.e. 10 euros.

## 4.2 Experiment 1: third party UG with an external proposer

In this experiment, we administered a version of the UG in which a *proposer* (P) allocates 10 euros between a *responder* (B) and a counterpart A. B could accept or reject the offer. P was not affected by B's decision; instead A receives the money offered by P if B accepted. The participant in the experiment plays the role of the responder (B). To control for the self-involvement, B was asked to play both for herself and on behalf of an unknown third party (B') (see figure 1), namely the participant that is playing as the *responder* in the next experimental session (see chapter two). The fact that an external person P decided the allocations made it plausible to present B with offers which are unfair but advantageous for her.



*Figure 4.1.* Experiment 1-Target manipulation. P makes offers to B on how to divide 10 euros between B and A. B decides whether to accept or reject these offers: accepting, the money is divided between B and A as P has decided, whereas rejecting, neither B nor A get anything. P is not affected by B's decision, and rejections lose their role of punishment. In a second condition, B is asked to decide on behalf of B': if B accepts P's offers, the money is divided between B' and A' as P has decided, otherwise neither B' nor A'



get anything. B' and A' are two (anonymous) participants that are playing in the roles of B and A in the next experimental session.

#### *4.2.1 Method*

##### *Participants.*

Twenty-eight healthy participants were tested (9 females), aged between 19 and 35. They were all paid the same fixed amount (10 euros), although in the beginning of the experiment they were told that their payoff would be dependent on their performance (see the paragraph “Overview on the experimental method”).

##### *Materials and procedure.*

The participant (B) is asked to accept or reject divisions of 10 euros between herself and another person A, who has no decisional power. Allocations are 9 (1:9, 2:8...9:1), and are made by an external proposer P (actually, they have already been established by the experimenter). Before starting the game, the participant reads the instructions sheet, in which she is told that three groups of participants takes part into the study: one group has played before as P, another is playing as A, and the last one plays as B, which is the group she has been randomly assigned to. The triplets A-B-P are randomly matched. The participant is told that P is a person that has already performed the task, and has received 6 euros for splitting several banknotes of 10 euros between two couples of players (A and B, A' and B'). The participant is also told that A is performing a memory task in another lab, and that she is completely unaware that her compensation depends on other persons' decisions. In this way, complete anonymity is guaranteed.

As in the previous studies described here, the participant is asked to accept or reject P's allocations also on behalf of an unknown third party, namely the next participant (B'); the

participant knows that B' is playing after her, and will decide on allocations made by another P between herself (B') and A', who is taking part to the same memory experiment as A does.

Following the same rule, the participant knows that she and A have been the B' and A' for the previous couple of participants. The participant knows that the payoff is a percentage reflecting the proportion of money she decided to accept. She knows also that B' (and, consequently, A') is actually given the percentage accepted on her behalf, thus B' starts playing with an endowment. In the same way, participant's payoff depends also on what the previous B has decided on her behalf.

In this setting, rejections lose their role of tools for punishing unfair intentions, since they do not affect the payoff of the potential unfair proposer (P). Moreover, participants can face unfair but advantageous divisions (e.g. 9 to B and 1 to A). Eventually, the self-involvement is controlled.

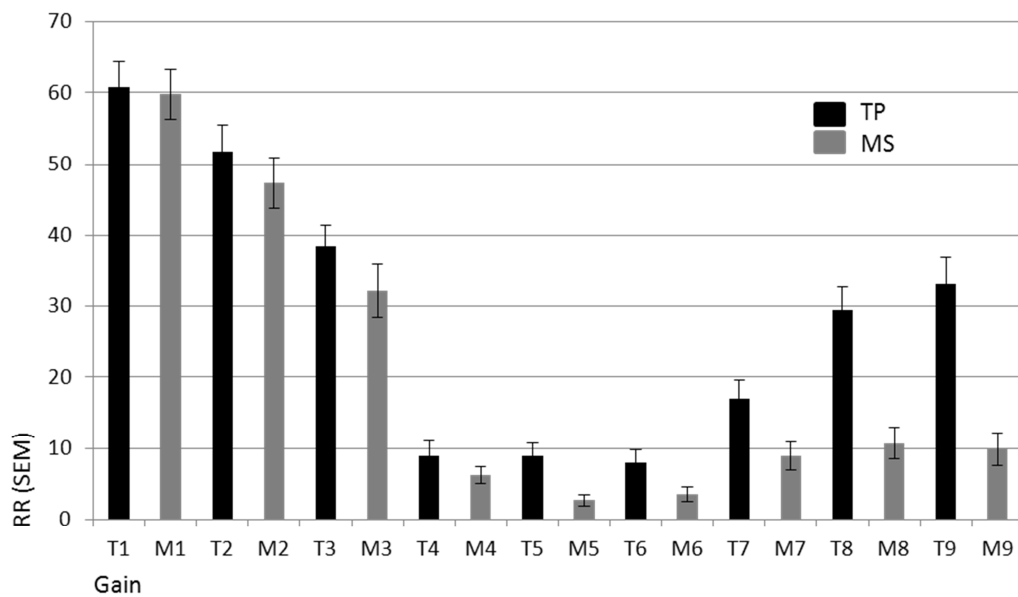
We have investigated the effects of two factors, which are Target of the division (2 levels: Myself -MS- and third party -TP-) and Gain (9 levels: 1 to 9 euros out of 10), on the rejection rate (RR). Each offer was randomly presented eight times (four in MS -myself condition: participant plays for himself- and four in TP -third party condition: participant plays on behalf of B'-) for 5 seconds, as follows: “P offers 7 euros to you (or “to the next participant”), and 3 euros to A (or “A’”)”. After 3 seconds, participants had to respond to the question “Do you accept?” by button press.

#### *4.2.2 Results and discussion.*

A repeated measures ANOVA (2X9) was performed, which shows a main effect of Gain ( $F(8, 208) = 26,965, p < .001, \eta_p^2 = .5$ ): RR is higher for disadvantageous offers and decreases as the offers become more advantageous, independently of the fairness.

Both the main effect of Target ( $F(1, 27) = 7,162, p < .05, \eta_p^2 = .210$ ) and the Target X Gain interaction ( $F(8, 216) = 1,998, p < .05, \eta_p^2 = .06$ ) were found to be significant: the RR is higher in TP for unfair advantageous offers with respect to MS (Figure 4.2 and Table 4.1).

However, RR in TP1 and in TP2 is higher compared to, respectively, TP8 and TP9 (TP1-TP9:  $t(27) = 3,974, p < .001$ ;  $t(27) = 3,101, p < .005$ ).



*Figure 4.2.* Results of experiment 1 (RR). The graph displays the mean percentage of offers rejected (y) for each gain (x). The error bars indicate the standard error of the mean (see Table 1). The significant main effect of Gain is given by the higher percentage of rejections for unfair disadvantageous gains (1,2,3) compared to fair (4,5,6) and unfair advantageous (7,8,9). The significant main effect of Target and the significant interaction TargetXGain are given by the higher percentage of rejections for TP8 and TP9

compared to MS8 and MS9. However, RR is higher for TP1 and TP2 compared to, respectively, TP8 and TP9.

*Table 4.1. Average RR for the disadvantageous (1+2+3+4), fair (5) and advantageous (6+7+8+9) offers, for MS and TP, respectively.*

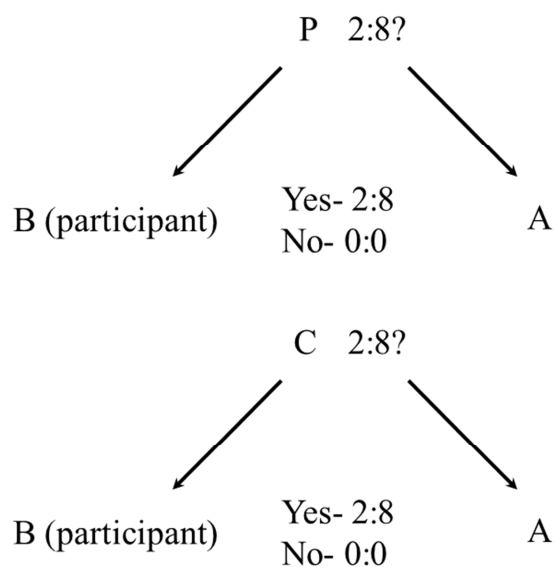
RR	DISADV	FAIR	ADV
RR_MS (SEM)	36.38 (6.02)	2.68 (1.49)	8.26 (3.77)
RR_TP (SEM)	39.95 (6.27)	8.93 (3.68)	21.87 (5.63)

If rejections are used as a tool to punish unfair intentions, they should not occur in any of the treatments in the present experiment, because B's rejections are not affecting P's payoff, hence punishment is not operating. The results show that rejections cannot be interpreted as punishment, because participants reject even if this decision affects the payoff of an "innocent" person, such as A, and do not punish the unfair person (P). Also the behavior of the responder acting for a third party cannot be interpreted by assuming that the subject takes on the preferences of the person she is deciding for. This result might be attributed to an in-group effect: it is possible that participants feel to be part of the group of the *responders*, given the fact that they know their payoffs are affected also by previous players' decisions, leading them to perceive the third party (B') as part of this group (Zamir, 2001; Civali et al., 2010).

Let us now consider inequality aversion. What should one take the preferences of the responder to be? If the theorist is again free to choose the preferences that fit the data, then assuming that player B has inequality aversion with lower weights on own utility will predict a pattern of rejection that we have observed.

### 4.3 Experiment 2: the UG with allocators' manipulation

In experiment 1 we have asked participants to accept or reject divisions of 10 euros between themselves and A, when these allocations were decided by an external proposer P. However, one can argue that, since intentions matter (e.g., Blount, 1995), the perceived unfairness of P might have triggered frustration, and hence rejections are a reaction to this frustration. In order to understand whether this is the case, in experiment 2 we have manipulated the allocator: now participants know that, in one condition, the division is made by P and, in a second condition, it is made by C, a random number generator, that gives the same probability to each outcome. C cannot be, by definition, unfair, since it does not have intentions. In addition to the predictions already discussed in experiment 1, here we also predict that, if rejections are triggered by P's intentional unfairness, they should disappear when the allocator is C.



*Figure 4.3.* Experiment 2- Allocator manipulation. B decides whether to accept or reject offers on how to divide 10 euros between him-/herself and A. Accepting, the money is divided between B and A as it has been established, whereas rejecting, neither B nor A get anything. In one condition, offers are made by P,

who is not affected by B's decision; in a second condition, offers are made by C, a coin. In both conditions, rejections lose their role of punishment.

#### *4.3.1 Method*

##### *Participants.*

Thirty healthy participants were tested (20 females), aged between 19 and 35. They were all paid the same fixed amount, i.e. 10 euros, although in the beginning of the experiment they were told that their payoff would be dependent on their performance.

##### *Materials and procedure.*

As in experiment 1, the participant (B) is asked to accept or reject divisions of 10 euros between herself and another person A, who does not play any role. Divisions' options are 9 (1:9, 2:8...9:1), and are made by either a third person P, who is not affected by B's decisions, as in experiment one, or by a random number generator C, which, by definition, cannot be unfair (offers are actually always established a priori by the experimenter). To avoid overloading participants with information, we have ruled out the myself-third party manipulation, asking them to play just for themselves (See fig. 4.3).

We have investigated the effect of two factors, Allocator (2 levels: P and C) and Gain (9 levels: 1 to 9 euros out of 10), on the rejection rate (RR). Each offer is randomly presented eight times (four by P and four by C) for 5 seconds, as follows: "P offers/C allocates: 7 euros to you and 3 euros to A". After 3 seconds, the participant has to respond to the question "Do you accept?" by button press.

#### *4.3.2 Results and discussion.*

A repeated measures ANOVA (2X9) was performed, which shows a main effect of Gain ( $F(8, 232) = 27,01, p < .001, \eta_p^2 = .482$ ). RR is high for disadvantageous offers and decreases as the offers become more advantageous, independently of the fairness/equity. No effect of Allocator was found (Figure 4.4 and Table 4.2)

From the results of experiment 2, it can be concluded, once again, that rejections do not depend on the will to punish the source of inequality. People tend to reject unfair and disadvantageous partitions even if they are the result of a random division made by C, and not of an intentional act of unfairness. And, again, inequality is not always rejected; in fact, it is accepted when it is advantageous for the responder's payoff. These results are apparently in contrast with the idea that intentions associated with human actions matters: we will discuss our interpretation in the General Discussion paragraph.

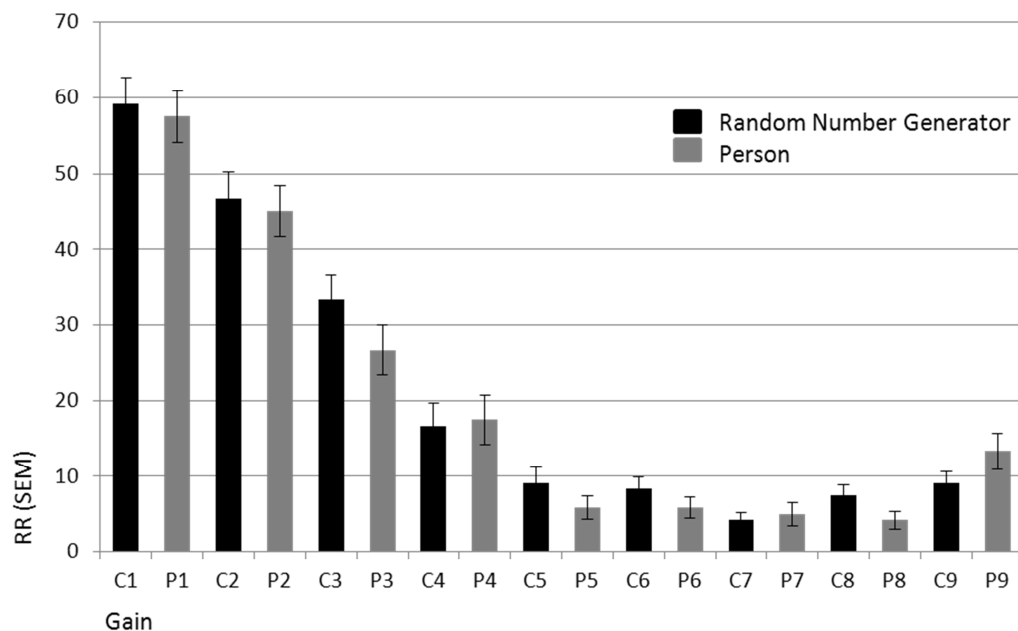


Figure 4.4. Results of experiment 2 (RR). The graph displays the mean percentage of offers rejected (y) for each gain (x). The error bars indicate the standard error of the mean (see table 4.2). The significant main effect of Gain is given by the higher percentage of rejections for unfair disadvantageous gains

(1,2,3) compared to fair (4,5,6) and unfair advantageous (7,8,9). No significant difference between the two allocators is found.

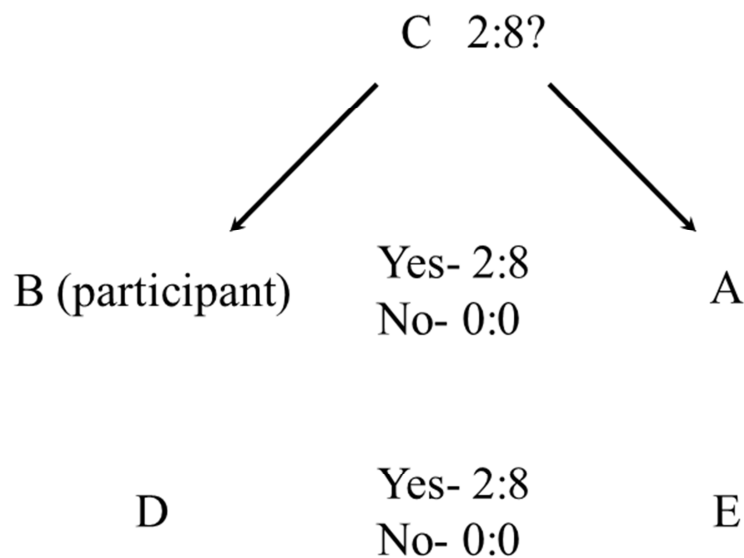
*Table 4.2. Average Rejection's Rate (RR) for the disadvantageous (1+2+3+4), fair (5) and advantageous (6+7+8+9) offers, for P and C, respectively.*

RR	DISADV	FAIR	ADV
RR_P (SEM)	36.66 (6.66)	5.83 (3.1)	7.08 (3.27)
RR_C (SEM)	38.96 (6.71)	9.16 (4.23)	7.29 (2.78)

#### **4.4 Experiment 3: third-party UG with coin's allocations**

In this next experiment, we kept the random number generator C as the allocator. We carried out this third experiment because we aimed at clarifying the third party behavior, namely the behavior of participants when playing on behalf of the third party. We have ruled out the potential in-group bias by telling the participant that she has to accept or reject the division between two unknown persons (D and E), who are participating to another experiment, as A does. In this way, the participant is not required to decide on behalf of the next responder, who, as discussed above, might be considered part of the participant's group. In addition to rejection rate, we have also collected fairness ratings, on a 12-point Likert scale ranging from “very fair” to “very unfair”. We predict that, if participants are endowed with a preference for fairness, they will reject all the unequal splits when the payoffs of D and E are involved, whereas they will reject only unequal split which are disadvantageous for themselves when their own payoff is involved, since in this case the selfish preference for a relative higher payoff may overcome the fairness preference.





*Figure 4.5.* Experiment 3-Target manipulation. C randomly establishes how to divide 10 euros between B and A. B decides whether to accept or reject these offers: accepting, the money is divided between B and A as C has established, whereas rejecting, neither B nor A get anything. Rejections lose their role of punishment. In a second condition, B is asked to decide whether to accept or reject divisions between D and E, who are two subjects completely uninvolved in the experimental protocol.

#### *4.4.1 Method*

##### *Participants.*

Forty-one right-handed healthy participants were tested (25 females), aged between 19 and 35. They were all paid the same fixed amount, although in the beginning of the experiment they were told that their payoff would depend on their performance.

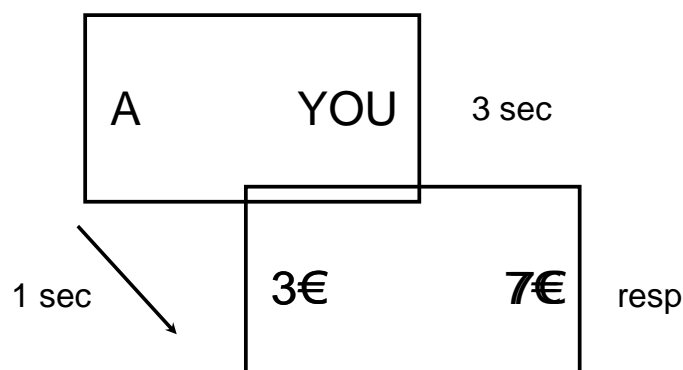
##### *Materials and procedure.*

As in experiments 1 and 2, the participant (B) is asked to accept or reject divisions of 10 euros between himself and another person A, who does not play any role. Divisions' options are

9 (1:9, 2:8...9:1), and are made by a random number generator C. The participant is also asked to decide on allocations between two third parties, not involved in the game; she is told that D and E (See fig. 4.5) take part into another experimental protocol, in which they are administered a memory task, and both of them are totally unaware of the way in which their payoffs will be decided, as A is.

We have investigated the effect of two factors, Target (2 levels: MS and TP) and Gain (9 levels: 1 to 9 euros out of 10) on the rejection rate (RR), while on fairness ratings the factors Target (2 levels: MS and TP) and Fairness (5 levels: unfair\_advantageous, mid-value\_advantageous, fair, mid-value\_disadvantageous, unfair\_disadvantageous) are considered.

Each offer is randomly presented eight times (four in MS and four in TP). The participant sees on the screen one couple of targets (“A, YOU” or “D, E”) (see fig. 4.6) for 5 seconds; after one second of blank screen, it appears the allocation (e.g. 3-7). The participant knows that the number on the left refers to the target previously presented on the left, and vice-versa. As soon as she sees the allocation, she has to respond by button press.



*Figure 4.6.* Experimental procedure. Participants saw the targets of the division (A, You / D, E) for 5 seconds on the screen. The position of each target is counterbalanced among the offers (e.g. for 3-7

division, which is repeated 4 times in MS, “You” appears twice on the left and twice on the right). After one second of blank screen, the division is displayed: participants know that, in this case, 3 refers to A’s payoff and 7 refers to their payoff. As soon as the allocation appears on the screen, participants have to respond “yes” or “no” by button press. Response button are counterbalanced among participants.

#### 4.4.2 Results and discussion.

The ANOVA on RR shows a significant main effect of Gain ( $F(8, 320) = 47.23, p < .001, \eta_p^2 = .541$ ), a significant main effect of Target (Target:  $F(1, 40) = 12.60, p < .001, \eta_p^2 = .240$ ) and a significant TargetXGain interaction ( $F(8, 320) = 19.345, p < .001, \eta_p^2 = .326$ ). In MS, as in the previous two experiments, RR is high for disadvantageous offers and decreases as the offers become more advantageous, independently of the equity; in TP, on the other hand, RR is higher for extremely unequal than for equal divisions (Figure 4.7). In contrast with experiment one, no significant difference was found for TP1-TP9 or TP2-TP8.

A repeated-measure ANOVA (2X5) was performed also on fairness ratings: it shows a main effect of Fairness ( $F(4, 156) = 99.906, p < .001, \eta_p^2 = .719$ ) and a significant interaction TargetXFairness ( $F(4, 156) = 3.874, p < .01, \eta_p^2 = .09$ ). The interesting result is the interaction; although MS unfair offers, both advantageous and disadvantageous, are still considered to be actually unfair, a t-test for paired samples shows that participants considered MS unfair\_advantageous offers significantly fairer than TP unfair offers ( $t(1, 40) = -2.599, p < .05$ ). From the results of experiment three, it can be concluded, once again, that rejections are not necessarily a punishing act, and that inequality is not always rejected; in fact, it is accepted when it is advantageous for the responder’s payoff. However, when deciding for third parties, free from potential in-group biases, participant has no reason to accept an unequal division, therefore inequality is generally rejected.

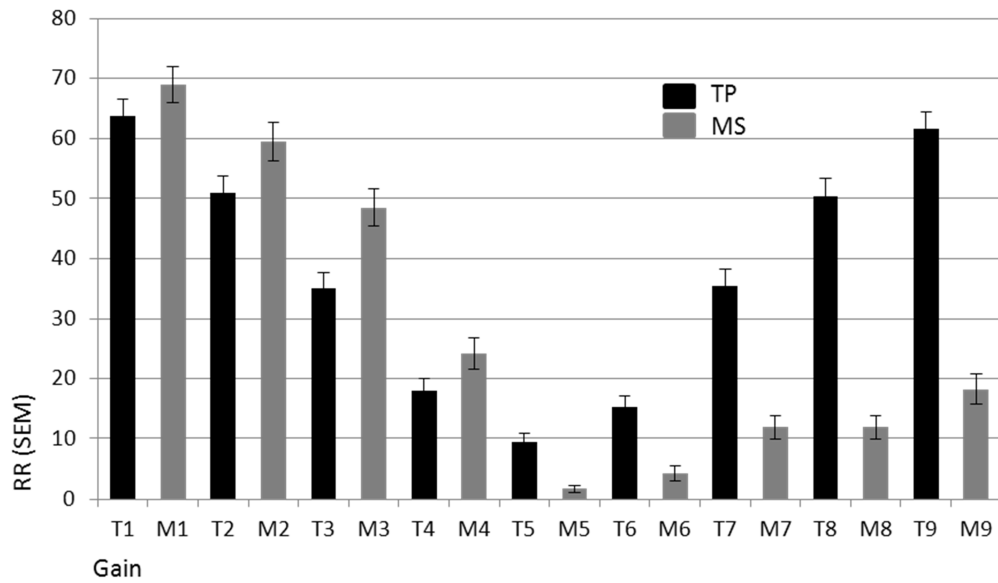


Figure 4.7. Results of experiment 3 (RR). The graph displays the mean percentage of offers rejected (y) for each gain (x). The error bars indicate the standard error of the mean (see table 4.3). The significant main effect of Gain is given by the higher percentage of rejections for unfair disadvantageous gains (1,2,3) compared to fair (4,5,6) and unfair advantageous (7,8,9) in MS. The significant effect of Target and the significant interaction TargetXGain is driven by the higher percentage of rejected offers for unfair advantageous gains in TP than in MS. The difference between the extremely unfair offers (1, 2, 3 and 7, 8, 9) in TP is not significant.

Table 4.3 Average Rejection's Rate (RR) for the disadvantageous/unfair (1+2+3+4), fair (5) and advantageous/unfair (6+7+8+9) offers, for MS and TP, respectively.

RR	DISADV/UNF	FAIR	ADV/UNF
RR_MS (SEM)	50.23 (5.88)	1.68 (1.21)	11.58 (3.87)
RR_TP (SEM)	41.92 (5.15)	9.45 (2.87)	40.62 (5.32)

## 4.5 General discussion

The traditional UG involves two players: the proposer, who makes offers on how to split an amount of money, and the responder, who accepts or rejects these offers. If the responder accepts, the money is divided as the proposer has decided, otherwise they both get nothing. Against predictions of self-interest made by classical economic theories, the proposer tends to make fair offers and the responder tends to reject offers that are considered unfair. The most accredited accounts that explain this apparently irrational behavior involve the concept of social preferences: individuals do not only care about their own payoff, but also about others' payoffs. In this view, rejections are seen both as a reaction to an unequal disliked outcome (Bolton and Ockenfels, 2000; Fehr & Schmidt, 1999) and as a punishment towards an unfair proposer (Rabin, 1993; Falk & Fischbacher, 2006). The wounded pride/spite model (Pillutla & Murnighan, 1996) recognizes in negative emotions, such as anger and frustration, the psychological cause of rejections: fairness norms are violated, anger and frustration are elicited by the violation, and individuals are driven to rejecting money irrationally in order to punish this violation.

However, the traditional UG has some limitations that do not allow understanding the roles of neither reciprocity, nor inequality aversion, nor negative emotions, in cause of the fact that a) rejections always lead to punish the source of unfairness, b) responders seldom face unfair but advantageous offers, and c) responders' payoff is heavily involved.

In the experiments described above, we have manipulated three variables (the involvement of the proposer, the advantage of the offers and the involvement of the self) in order to clarify what influences choices in this bargaining game.

### 4.5.1 *Negative reciprocity*

A robust and crucial result is that negative reciprocity does not seem to be the sole cause of the rejections, thus leading to reject our second hypothesis; our responders rejected offers even if it did not mean to punish unfair behavior, and this occurred both when responders believed the allocator was an external, non-involved proposer, and when they believed the allocation was set by a random number generator. This finding seems to contrast theories of reciprocity, which claim that rejections are driven by the desire to punish the bad intentions of the proposer. However, there are a few points that differentiate the current study from the previous ones. First of all, I will consider the case of the reduced UG (Falk, Fehr, and Fischbacher, 2003; Güroğlu, van den Bos, Rombouts, and Crone, 2008). In this version of the game, proposers have to divide 10 dollars, and have to choose between two alternatives: in condition a), one alternative is fair (5/5) and the other is unfair (8/2 for the proposer); in condition b), one alternative is unfair (8/2 for the proposer) and the other is hyper-unfair (10/0 for the proposer). In a), responders reject the unfair offer, while in b) they accept it, even though it is exactly the same offer. This result is considered to be a proof of the importance of the proposers' intentions over the simple consideration of the outcome. However, whereas in this study responder's attention is focused on the proposer's intentions, giving her different choices, in the current experiments participants were not focused on the intentions: in fact, they just could not attribute any kind of responsibility, and this might have ruled out any kind of considerations of good or bad intentions from the decisional process.

Secondly, I am considering the study of Blount (1995). Here, the author found that when the allocations were decided by a random number generator, people tend to state lower MAO, compared to the condition in which the division was set by the proposer. I account this difference to the different requirements of the task. In Blount's experiment, participants were first asked to predict a distribution of offers, and then to decide a MAO; requiring participants to focus on the

distribution, which, in the random condition, was expected to be flat and to place equal probabilities across all possible outcomes, might have influenced the subsequent responses by driving participants to make MAOs more consistent with the idea of the random distribution. However, still the mean MAO is 1.20 \$, indicating that offers which are considered too low would be rejected anyway. In our task, participants are directly asked to accept or reject the offers, and so they do not have the chance to focus on what they would rationally expect from the game. This leaves participants to be more instinctive in their choices, and free from expectations biases (Sanfey, 2009). Moreover, in Blount's study, there is no difference between the condition in which the allocation is decided by the involved proposer and the condition in which it is decided by an external proposer. Even if intentions are still considered, reciprocity should not be accounted for this result; as Blount noted, responders can be motivated either by a desire to manifest their disappointment and exit the game if the procedure produces unfair results or by desiring to punish the proposer who behaved unfairly. We are not claiming here that the negative reciprocity does not play any role in this interaction: in fact, it might be that it is the main reason to reject unfair offers made by an involved proposer (traditional UG). However, since people reject also when negative reciprocity cannot be accounted for the rejections, it means that there are different factors involved. Though we cannot say if in the traditional UG prevails one or the other motive, or if they coexist, we can claim that, given our results, the desire to refuse general unfairness, beyond punishment, exists and plays a crucial role under certain environmental conditions.

#### *4.5.2 Inequality aversion.*

What few studies, to my knowledge, have focused on is the maximum acceptable offer that responders are willing to accept (Mitzkewitz and Nagel, 1993). There are studies on social values orientation that divides people into three categories, which are prosocials, individualists and competitors, depending on which aspect of the division the person cares about. Prosocials are interested in increasing their payoff together with the payoff of the counterpart, trying to minimize the discrepancies; individualists care only about maximizing their own gain; competitors seek, in contrast to prosocials, to maximize discrepancies between relative payoffs (Van Lange, 1999; Haruno & Frith, 2010). These studies used the Ring Measure of Social Value to assign people to one of these three categories (Liebrand, Jansen, Rijken, and Suhre, 1986): in this paradigm, the person distributes hypothetical amounts of money between himself and another person, choosing the distribution between two alternatives. However, there are limitations to this task that should be considered: first of all, the money is hypothetical, and, in van Lange's study, participants knew since the beginning that their payoff was not influenced by their performance, thus limiting the expression of the social preference to a hypothetical context; secondly, the reputation effects are not controlled, and it is known that these are variables that can influence dramatically the behavior, as it is demonstrated in Hoffman, McCabe, Shachat, and Smith (1994) and in Dana, Kuang, and Weber, (2007). In conclusion, both the limit of hypothetical scenario and the reputation-seeking effects can induce the participants to show a prosocial behavior that actually depends on self-interested motivations concerning reputation.

In my studies, I have ruled out the reputation effects by guaranteeing the total anonymity; I have used a modified UG, giving the responders the opportunity to accept or reject unfair but advantageous offers, finding out that they were willing to accept them, even if they still



considered them to be unfair, as demonstrated by fairness ratings results in experiment three. For this reason, I question the account that considers the rejections to be caused by a taste for fairness and an aversion for unequal payoff. One of the most accredited accounts that describe inequality-aversion is the one put forward by Fehr and Schmidt (1999): as mentioned in the introduction, the authors predict that a responder is inequality averse, but with a preference for accepting unequal outcomes that favor her own payoff. The results presented here for the ms condition would actually confirm their theory of inequality aversion: responders reject unfair offers, but accept as the offers become advantageous for themselves. However, I am skeptical in considering this as an evidence of social preference for fairness: if a preference for fair outcomes implies the rejection of disadvantageous unfair outcome but the acceptance of advantageous unfair outcomes, it actually turns to be a preference for being better off the other player. Fehr and Schmidt consider two parameters, one representing the degree of fairness concerns and the other one representing the envy, and they are actually able to explain most of the behavioral evidences varying these parameters; however, as Bicchieri and Zhang (2008) have point out, they do not specify how these parameters should vary, and which are the circumstances that might explain the variance.

As far as the third party condition is concerned, the theories of inequality aversion make no specific prediction: in fact, Fehr and Schmidt say “inequity aversion is self-centered if people do not care per se about inequity that exists among other people, but are only interested in the fairness of their own material payoff relative to the payoff of others” (Fehr & Schmidt, 1999, p.819). On the contrary, my results show that people seem to care about inequality that exists among other people, reflecting a preference for fair outcomes, as demonstrated by the fact that unfair outcomes in tp are rejected. Therefore, it could be argued that people are endowed with

fairness concerns, which turn into more selfish self-advantage considerations when their own payoff is at stake.

#### *4.5.3 Self-involvement*

In experiments 1 and 3, as well as in the studies reported in chapters 2 and 3, I have manipulated the degree of personal involvement of the responder. This was necessary in order to understand whether rejections are driven by an irrational reaction to a personal attack, or by a more “cognitive” moral aversion to fairness' norm's violation. The results presented here showed that the degree of self-involvement plays a crucial role: when the responder's payoff is at stake, a more competitive behavior overcomes fairness concerns, which are taken into account, in contrast, when participants play on behalf of third parties. In support to this interpretation, I have found a difference in the rejection rate by manipulating the degree of closeness of the third party involved: in experiment 1, when the third party is depicted as the next responder and thus considered as part of the participant's group, the rejection rate for (unfair) offers which are advantageous for the third party is lower than the rejection rate for (unfair) disadvantageous offers, while we have seen no difference for rejection rate in the uncorrelated third party condition (experiment 3), in which participants accepted only fair offers. This can be explained in terms of favoritism: responders tend to favor themselves and their group members, whereas when there is no involvement, and no reason to favor one person over the other, they are inequality-averse. This observation might sound trivial, but actually self-involvement is something that has always been neglected when considering social preferences. This happens because, classically, self-involvement is an integral part of the game; however, theories on social preferences should take into account the strength of this variable, and consider that basic

behavioral principles might change just according to the level of involvement. This is a phenomenon that can be experienced every day: for instance, we know that we are very good in judging people as far as we, or our closest ones, are not those people.

In light of these results, I propose an account in which three principles are considered to explain responder's behavior: maximization of the payoff (rationality), competitiveness (self-oriented fairness) and pure fairness preference. When the responder is personally involved in the bargain, the three principles vie to the response; when she is not personally involved, rationality principle and fairness are competing for the response. Therefore, I am concluding that deciding on monetary outcomes depends on preferences associated with the material payoff; however, preferences vary with contextual cues, such as the self-involvement, which determines a shift of tolerance for unequal outcome towards self-advantage. Social preferences are not stable; theoretical accounts that consider environmental changes as predictors of behavior are the best candidates to explain behavioral dynamics that take place in social games. (Dana et al., 2007; Bicchieri & Zhang, 2008; Bicchieri and Xiao, 2008; Hertwig & Herzog, 2009). Gneezy and Rustichini, for example, found that participants who were rewarded with a monetary compensation for their performance in a cognitive task performed less successfully as compared to participants who were just driven by the intrinsic motivation, which is in itself rewarding (Gneezy & Rustichini, 2000); this means that the utility is not always to be considered as the monetary reward, but, depending on the situation, it might take different shapes.

In this study, we have shed light on different aspects of social interaction in bargaining games, showing that the principles that drive the behavior might change together with self-involvement. We have seen that pure inequality-aversion and fairness concerns are leading behavior when the participants are not directly involved, whereas maximizing the differences

between payoffs seems to be the major concern when participants' own payoff is at stake. These findings support the account according to which behavioral responses are guided by social strategies that are, in turn, triggered by environmental changes. This assumption is plausible if it is considered that our cognitive system seems to be accustomed to strategies. to take an example from cognitive neuroscience, it has recently been shown that, in contrast to one of the main assumptions of embodied cognition theories, the motor system is not necessarily required in order to understand motor-related language and that its recruitment is just strategic (Tomasino, Werner, Weiss, and Gereon R. Fink, 2007; Tomasino, Fink, Sparing, Dafotakis, and Weiss 2008; Papeo, Vallesi, Isaja, and Rumiati, 2009; Papeo, Corradi-Dell'acqua, and Rumiati, 2011). If the context does not allow this recruitment (e.g., lesions, interferences), other paths are chosen in order to achieve the task.

Further work is needed; in particular, also in light of results presented in chapter 3, we hypothesize that the self-centered inequality aversion might be grounded in self-related and emotion-related areas, such as the medial prefrontal cortex whereas the pure fairness concerns might have their neural correlates in areas associated to morality and norm compliance, such as anterior insula.

## Chapter 5

# The neural basis of inequality as an abstract social rule.

### 5.1 Introduction

In chapter three, I have discussed the involvement of the medial prefrontal cortex (MPFC) and the anterior insula (AI) in UG rejections; MPFC was involved in rejecting unfair offers which addressed the self (MS condition), whereas AI in rejecting unfair offers that addressed both the self and the third party (TP condition). The data suggest that this middle-anterior MPFC activity might be related to emotional arousal evoked by an unfair offer related to the self, while AI seems an ideal candidate for mediating *fairness*-related behavior which emerges from the integration of cognitive, emotional and motivational mechanisms. However, it is still unclear whether the *fairness*-related behavior reflects a moral act, motivated by the wish of sanctioning an intentional unfair action, or inequity aversion, motivated by the wish of preventing an unfair division from taking place, irrespectively of its moral salience. Moreover, it has also to be clarified whether MPFC and AI respond to unfairness itself or to unfairness which is

disadvantageous for the responder; in fact, the fMRI study described in chapter 3, even though with the introduction of the TP condition made it possible to dissociate general unfairness from unfairness directed to the self, does not really allow to disentangle between norm compliance and empathy in the TP condition, given the possible identification of the responder with the next responder (in-group effect).

In the last chapter, I have described data supporting the idea that pure inequity-aversion and fairness concerns are leading behavior when the participants are not directly involved, whereas maximizing the differences between payoffs seems to be the major concern when participants' own payoff is at stake. Moreover, I have found no difference in the rejection rate between a human non-involved proposer and a partition algorithm (see experiment 2, chapter 4). This means that, first of all, even though bad intention may amplify the reaction to unfairness, subjects are sensitive to inequity. Secondly, as Fehr and Schmidt (1999) predicted, inequity-aversion is actually self-centered, since responders are more averse to proposals that favor the opponent, and more prone to accept divisions that favor themselves. This is true when the self is involved, while it is not the case in the TP condition, when a completely unrelated third party is involved: in this situation, the default mode is rejecting unfairness, when there is no good motivation to accept it, such as merit or need (e.g. Camerer & Thaler, 1995).

In order to contribute to the comprehension of the brain mechanisms underlying this complex behavior, I have carried out an fMRI study in which subjects, while lying in the scanner, played as responders in a task which was very similar to the one described in the third experiment of chapter 4: participants had to accept or reject divisions made by an algorithm, both between themselves and an uninvolved opponent A, and between two unrelated third parties, E and D. The hypothesis is that the self-centered inequity aversion might be grounded in the

emotional system, whereas the pure fairness concerns might have their neural correlates in areas associated to morality and norm compliance. I expect a major activation in the MPFC during unfairness in the MS condition with respect to the TP condition, especially for disadvantageous unfairness; consequently, I expect also a major activation when rejecting unfair offers, as opposed to acceptance, in MS, but not in TP. Also, AI is expected to be involved in facing unfairness, irrespectively of whether this is advantageous or disadvantageous, or self- or other-directed.

## **5.2 Methods**

### *5.2.1 Participants*

Nineteen (12 females) subjects took part in the experiment. None of the participants had any history of neurological or psychiatric illness. Written informed consent was obtained from all subjects, who were naive to the purpose of the experiment. The study was approved by the local ethics committee.

### *5.2.2 Task and Stimuli*

Task, stimuli and experimental set-up were similar to those used in experiment three as described in chapter 4.

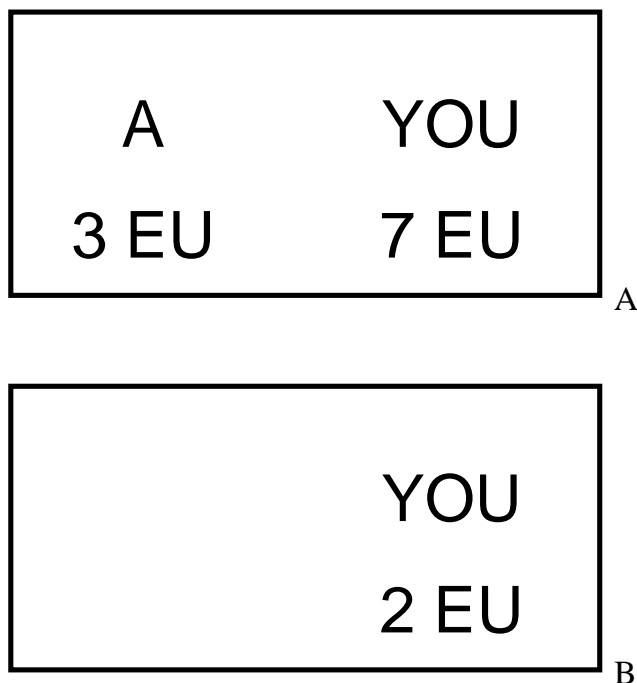
To sum up briefly, the random number generator C is kept as the allocator. The potential in-group bias has been ruled out by telling the participant that she has to accept or reject the division between two unknown persons (D and E), who are participating to another experiment, as A does. In this way, the participant is not required to decide on behalf of the next responder, who, as discussed above, might be considered part of the participant's group. As in the previous

experiment, I predict that, if participants are endowed with a preference for fairness, they will reject all the unequal splits when the payoffs of D and E are involved, whereas they will reject only unequal split which are disadvantageous for themselves when their own payoff is involved, since in this case the selfish preference for a relative higher payoff may overcome the fairness preference (see figure 4. 5, chapter 4, for the structure of the task). In this experiment, I have re-introduced the Free Win (FW) task, in which participants have to accept or reject money provided by the computer that is not resulting from a partition, in order to control for pure monetary gain cleaned from fairness effects. As opposed to both the physiological study described in chapter 2 and imaging study in chapter 3, this time the FW task refers only to the myself condition, not to the third party condition.

I have investigated the effect of two factors, TASK (3 levels: UG\_MS, UG\_TP and FW) and GAIN (5 levels), on both the behavior and the brain activity. Division options were 9 (1:9, 2:8....8:2, 9:1), repeated 4 times each except from 5:5, which was repeated 8 times; divisions were consequently collapsed in hyper\_disadvantageous –h\_dis- (1:9, 2:8), mid\_disadvantageous –m\_dis- (3:7, 4:6), hyper\_advantageous –h\_adv- (9:1, 8:2), mid\_advantageous –m\_adv- (7:3, 6:4) and fair –for this, the factor GAIN has 5 levels-. For FW, the offers range from 1 to 9 (for an example of both UG and FW trials, see figure 5.1, A and B respectively). It is necessary to stress that this terminology is suitable for the myself condition, in which there is actually a clear advantage or disadvantage for the subject, whereas it is not properly used for the third party condition, where it would be better to consider the offers as two groups of hyper\_unfair –h\_unf- and two groups of mid\_unfair (one group favors C and the other favors D), as the subject is getting



advantage (or disadvantage) from none of them. Also, as far as the FW is concerned, terms related to fairness have no real meaning: in this task no fairness should be involved, since no comparison is going on. For simplicity, I will refer to h\_dis, mid\_dis, mid\_adv, h\_adv and fair\_ms when referring to UG\_MS and when describing the model in general, and to h\_unf, mid\_unf and fair\_tp when referring to UG\_TP.



*Figure 5.1. (A)* An example of one UG\_MS trial: it lasted 3 seconds on the screen, and the participant had to respond by button press as soon as the offer appeared on the screen. The position of the targets “A” and “YOU” was counterbalanced among trials, in order to rule out potential effects given by attention, compatibility or number line. The same structure applies for UG\_TP trials, with D and E instead of A and YOU. *(B)* An example of one FW trial: also in this case, the position of the target was counterbalanced among trials.

### *5.2.3 fMRI data acquisition and experimental set-up*

Images were acquired using a 3-T MRI scanner (Achieva 3.0T Philips Medical Systems, Netherlands) equipped with a standard quadrature head coil and for echo-planar (EPI). Head movement was minimized by mild restraint and cushioning. Thirty-four slices of functional MR images were acquired using blood oxygenation level-dependent (3.59 x 3.59 mm, 4 mm thick, repetition time = 2 s, time echo = 35 ms; flip angle: 90; field of view, FOV: 23 x 23 cm, acquisition matrix: 64x64; SENSE factors: 2 in anterior-posterior direction), covering the entire cortex. At the beginning of the scanning session, anatomical scans were also acquired for each participant (TR/TE: 8.2/3.7, 190 transverse axial slices; flip angle: 8; 1 mm<sup>3</sup> voxel size; FOV=24 cm x 24 cm; acquisition matrix: 240x240; no SENSE factors).

Participants lay supine in the MR scanner with their head fixated by firm foam pads for approximately 40 minutes (7 minutes required for the anatomical scans and about 32 minutes for the task). The experimental task was presented using the Presentation software (Neurobehavioral Systems, Inc.), and delivered within the scanner by means of MR-compatible goggles mounted on the coil. For each experimental trial, participants were first presented with the monetary offer for 3000 msec, to which they had to respond the faster they could by button press; trials were followed by an inter-trial interval ranging from 3000 msec to 7000 msec with an incremental step of 33 msec. Each experimental session comprised 120 randomized trials [3 TASK x 5 GAIN x 8 repetitions] and 1 minute of low level baseline, as 20 seconds of fixation cross at the beginning, in the middle and at the end of each run.

#### *5.2.4 Behavioral and imaging data processing*

For each subject, and for each condition, the rejection rate was calculated across all 8 repetitions, and used in a TASK X GAIN Repeated Measures ANOVA. Statistical analysis was performed using SPSS 11.5 Software (SPSS Inc., Chertsey UK).

Image processing and statistical analysis were performed using the SPM8 software package (<http://www.fil.ion.ucl.ac.uk/spm/>). For each participant we acquired 1030 volumes (515 volumes for each fMRI-run); the first 5 volumes were discarded for each run to allow the magnetization reach steady state. Slice-acquisition delays were corrected using the middle slice as reference. All images were corrected for head movements. All images were then normalized to the standard SPM8 EPI template and spatially smoothed using an 8 mm FWHM Gaussian filter. The high-pass filter was set to the cut-off value of 128 s.

Data were then fed into a first level analysis (GAIN model) using the general linear model framework (Kiebel and Holmes, 2004) implemented in SPM8. On the first level, for each individual subject, I fitted a linear regression model to the data. A factorial design 3 (TASK: UG\_MS, UG\_TP, FW) x 5 (GAIN: h\_dis, m\_dis, fair, m\_adv and h\_unf) yielded to 15 conditions. These 15 vectors were convolved with a canonical haemodynamic response function; to account for movement-related variance, we included six differential realignment parameters as regressors. The first level analysis of each subject yielded images describing the parameter estimates associated with each of the vectors modeled. These images were then fed into a second-level full factorial design with a within-subject factor of 15 levels using a random effects analysis (Penny and Holmes, 2004).

## 5.3 Results

### 5.3.1 Behavioral results

Overall, rejection rate is smaller with respect to the previous experiments presented, especially as far as MS condition is concerned; in fact, 6 subjects over 19 did not reject at all, other 3 subjects never rejected in MS, and one subject never rejected in TP. For each of the 19 subjects and for each condition, the rejection rates were calculated across all 8 repetitions, and used in a 3 TASK x 5 GAIN Repeated Measures ANOVA. Results indicate a significant main effect of TASK ( $F(2, 36) = 16.101, p < 0.001$ , part. eta squared=.472), with the UG\_TP leading to a larger number of rejections than the UG\_MS and the FW, as well as a main effect of GAIN ( $F(4,72) = 11.761, p < 0.001$ , part. eta squared=.395), with unfair offers being rejected more often than fair offers. These effects are driven by a TASK \* GAIN interaction, which is also significant ( $F(8,144) = 4.189, p < 0.001$ , part. eta squared=.189), suggesting that unfair offers in TP are rejected significantly more often than h\_adv ( $t(18)=-3.92, p<.001$ ), but not than h\_dis ( $t(18)=-1.72$ ), offers in UG\_MS (see Figure 5.2).

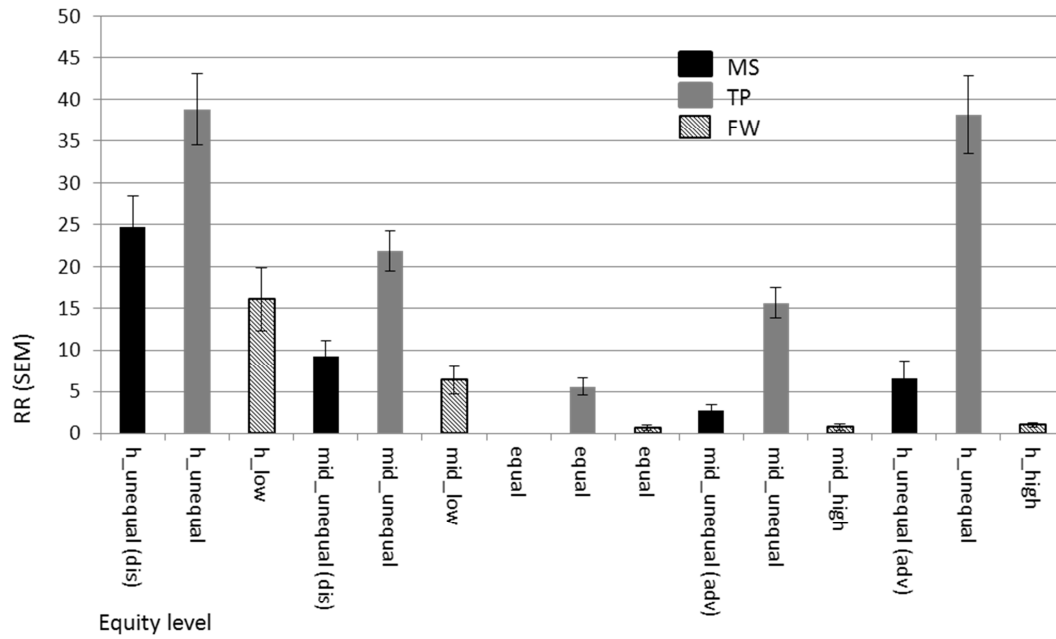


Figure 5.2. Behavioral Results. Rejection Rates are plotted as a function of Gain (Equity level) in UG\_MS, UG\_TP and FW. The error bars indicate the standard error of the mean (SEM) (see table 5.1)

Table 5.1. The table reports the average RR and the standard errors of the mean (SEM) for h\_dis, mid\_dis, fair, mid\_adv, h\_adv (columns), and for MS, TP and FW (rows).

RR	h_dis	mid_dis	fair	mid_adv	h_adv
RR_MS (SEM)	24.70 (7.48)	9.21 (3.77)	0	2.67 (1.47)	6.58 (4.12)
RR_TP (SEM)	38.79 (8.57)	21.84 (4.74)	5.59 (1.03)	15.61 (3.70)	38.15 (4.74)
RR_FW (SEM)	16.11 (7.59)	6.38 (1.63)	0.65 (0.65)	0.75 (0.75)	1 (0.55)

### 3.3.2 Neural Activations

Unless stated otherwise, we report exclusively areas of activation which survived threshold of  $p < 0.05$ , corrected for multiple comparisons at the cluster level, using FWE, with an

underlying height threshold of  $t > 3.12$  (= cluster size estimate to  $p < 0.001$ , uncorrected) (Table 5.1).

*Table 5.2.* Voxels showing significant increases of neural activity associated with main effects and interactions. All clusters survived a threshold corresponding to  $p < 0.05$ , corrected for multiple comparisons across the whole brain, with an underlying height threshold corresponding to  $p < 0.001$  (uncorrected). Only contrasts yielding significant activations are reported.

REGION	SIDE	MNI COORDINATES			Z-value	kE
Main effect of TASK:						
UG_MS > UG_TP		x	y	z		
Middle temporal gyrus	L	-56	-50	6	4.53	672
Superior temporal gyrus	L	-66	-46	12	3.70	
Main effect of UNFAIRNESS:						
Unf>Fair						
Inferior frontal gyrus	R	36	24	-8	5.66	933
Cingulate cortex	L	-4	18	46	5.05	1454
Anterior cingulate cortex	R	6	30	24	4.10	
Anterior cingulate cortex	L	-4	28	26	3.82	

Inferior frontal gyrus	L	-44	4	30	4.46	330
Anterior Insula	L	-38	14	4	4.45	537
<b>Simple main effect of UNFAIRNESS:</b>						
<b>UG_MS_unf&gt;UG_MS_fair</b>						
Inferior frontal gyrus	R	34	22	-14	4.83	661
Medial frontal gyrus	R	8	22	50	4.09	581
Superior frontal gyrus	R	12	26	60	3.95	
Anterior cingulate cortex	L	-8	24	32	3.26	
Inferior frontal gyrus	L	-34	16	-10	3.78	255
Anterior Insula	L	-32	20	8	3.46	
<b>Simple main effect of UNFAIRNESS:</b>						
<b>UG_TP_unf&gt;UG_TP_fair</b>						
Inferior frontal gyrus	R	34	26	-2	4.67	206
<b>Interaction:</b>						
<b>UG_MS (h_dis + mid_dis)- [UG_MS(h_adv+mid_adv)+UG_TP(h_unf+mid_unf)]</b>						
Superior frontal gyrus	R	14	52	40	4.76	797
Superior frontal gyrus	L	-8	56	40	4.11	

Medial frontal gyrus	R	10	42	40	3.64	
Supramarginal gyrus	L	-56	-68	30	4.03	598
Middle temporal gyrus	L	-54	-52	4	3.88	
Inferior parietal lobe	R	54	-60	38	3.90	384
Superior temporal gyrus	R	52	-60	28	3.87	

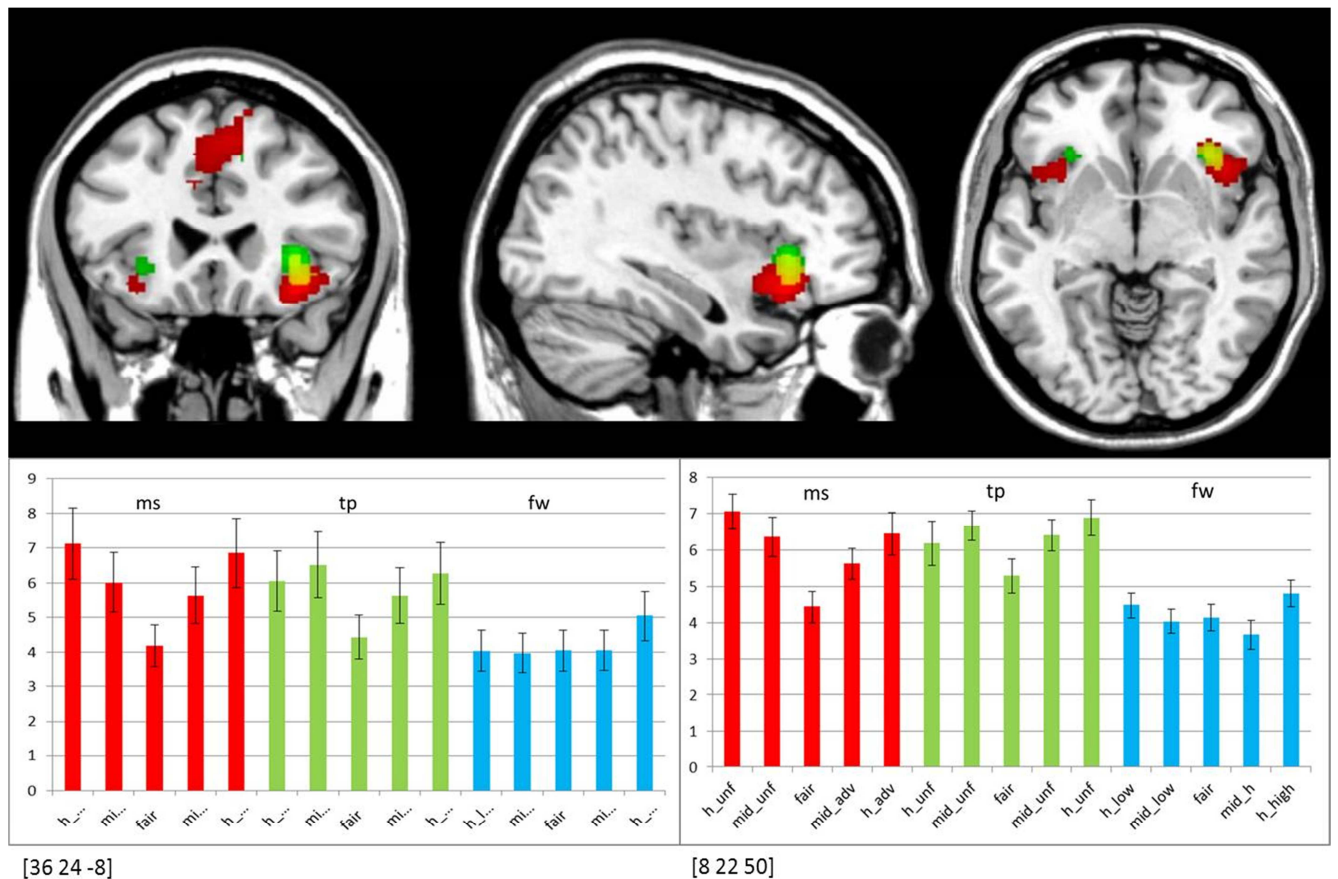
*Main effects.* I first tested for increases of neural activation associated with offers addressing oneself as opposed to offers addressing a third-party, in the UG [i.e., UG\_MS–UG\_TP]; such increase was found significant in the posterior superior and middle temporal gyrus. No suprathreshold voxel was found for the opposite contrast. Then, I looked for activations related to the unfair offers, as opposed to fair, across targets, i.e. UG\_MS [(h\_dis+mid\_dis+h\_adv+mid\_adv)-fair] + UG\_TP[(h\_unf+ mid\_unf)-fair]; a significant increase of BOLD signal was found in the left and the right AI and inferior frontal gyrus (IFG), and in the anterior cingulate cortex (ACC) extending more ventrally and more anteriorly in the medial prefrontal cortex (MPFC), bilaterally. No suprathreshold voxel was found for the opposite contrast. From the analysis of the two simple main effects of unfairness, considering separately UG\_MS from UG\_TP, it emerged that the activation in the MPFC was limited to UG\_MS, whereas the AI was activated in both tasks. As shown in the signal plots, both left and right AI and MPFC clearly showed a U-shaped modulation as a function of the level of unfairness; for this reason, quadratic contrasts were performed on UG\_MS, UG\_TP and FW, assigning different



weights to the different levels of fairness, i.e.  $[2 \ 1 \ -6 \ 1 \ 2]$ . A cluster involving inferior frontal gyrus and anterior insula bilaterally was significantly active for both the UG\_MS and in the UG\_TP conditions, although the cluster in the latter condition was smaller than the cluster in the former (see figure 5.4, red and green blobs for, respectively, MS and TP, and graph), whereas a cluster extending from the dorsal to the ventral medial prefrontal cortex, over and around the anterior cingulate cortex (figure 5.4 and figure 5.5, red blobs) was active specifically for UG\_MS. As expected, no suprathreshold voxel was found for the quadratic contrast in FW.

*Interactions.* A main effect of task (UG\_MS>UG\_TP) and a main effect of fairness (unf>fair) were found; in order to understand whether there were areas which were specifically activated by unfair offers, as opposed to fair, in UG\_MS, but not in UG\_TP, I performed a contrast which tested for the interaction between task and fairness, i.e. UG\_MS [(h\_dis+mid\_dis+h\_adv+mid\_adv)-fair] – UG\_TP [(h\_unf+mid\_unf)-fair]. No significant activation was found for our specific cluster level correction criterion; however, when cluster size was estimated at  $p=0.005$ , (instead of  $p=0.001$ ) a portion of the MPFC showed a significant activation. As shown in the signal plots, the trend of the activation was higher for the disadvantageous as opposed to the advantageous offers in MS. Thus, in order to investigate the peculiarities of brain activations in relation to disadvantageous offers, this condition was contrasted with all the other unfair offers, both in UG\_MS and in UG\_TP. As expected, the contrast UG\_MS (h\_dis + mid\_dis)- [UG\_MS(h\_adv+mid\_adv)+UG\_TP(h\_unf+mid\_unf)] showed an activation in the same cluster of the MPFC ([14 52 40),  $kE=797$ ), meaning that this area is more sensitive to the unfairness when it is self-disadvantageous as opposed to self-advantageous or other-affecting

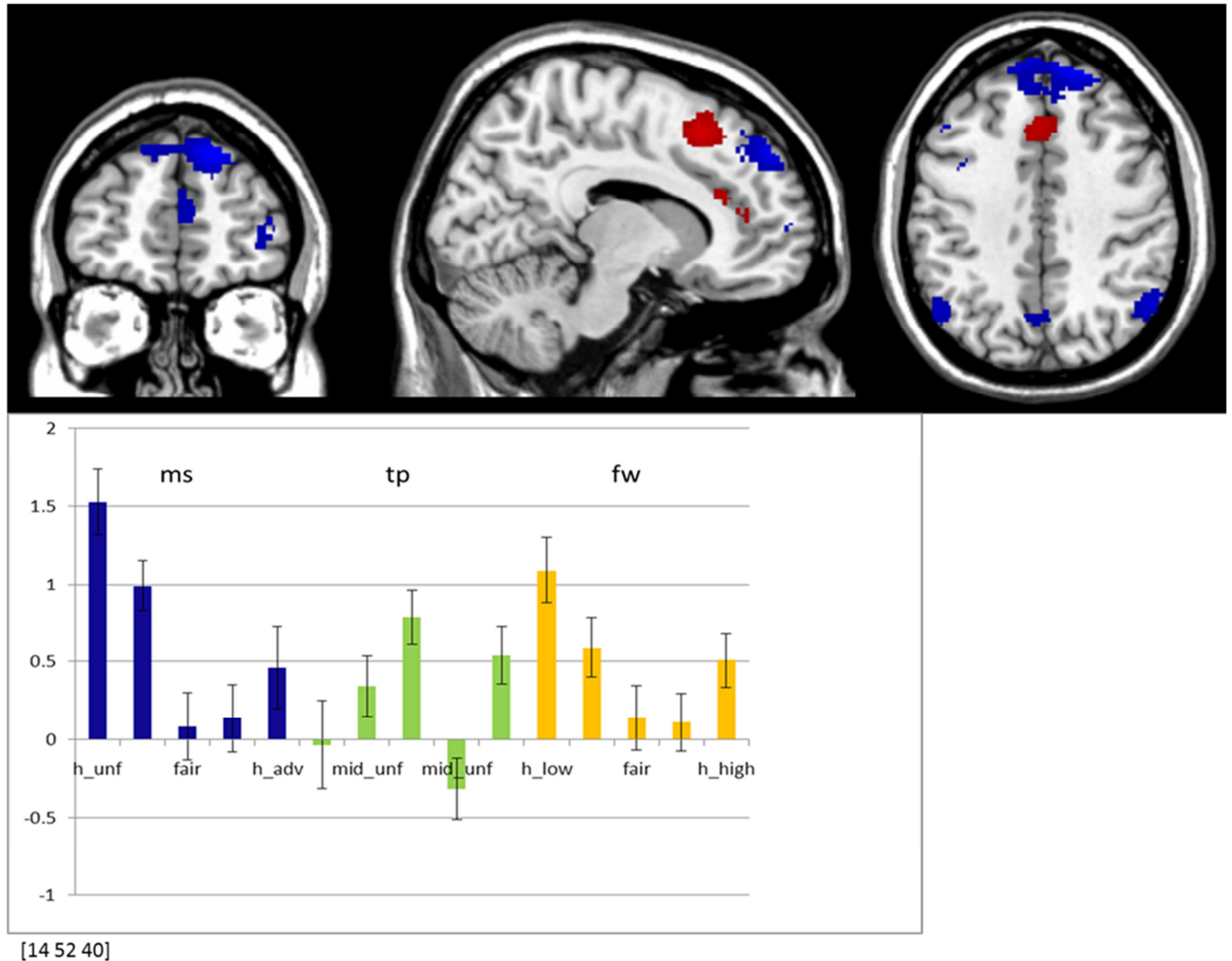
unfairness (see figure 5.5, blue blobs and graph).<sup>4</sup> No significant suprathreshold voxel was found for the opposite subtraction.



*Figure 5.4.* In the upper part, the activation for the main effect of unfairness (unf>fair) is depicted. The red blobs indicate the activation for the UG\_MS (MPFC+AI/IFG), the green blobs the activation for UG\_TP (AI/IFG). In the lower part of the figure, the graph on the left (x=fairness level; y=beta-values; red bars=MS; green bars=TP; blue bars=FW; error bars=SEM) shows the activation in AI/IFG [36 24 -8]; a clear U-shape trend is recognizable for UG\_MS and UG\_TP, confirmed by the quadratic contrasts, while there is no effect for FW. The graph on the right shows the activation of the MPFC [8 22 50], which

<sup>4</sup> Considering only UG\_MS, the same area was found significantly activated for disadvantageous, as opposed to advantageous offers, even though it survived a threshold of 0.005 instead of 0.001.

is visible on the coronal section (red blobs), and in figure 5.5; in this case, the parametrical modulation of fairness level (U-shape) is significant for UG\_MS, but neither for UG\_TP nor for FW.



*Figure 5.5* In the upper part, the activation for the interaction UG\_MS (h\_dis + mid\_dis)-[UG\_MS(h\_adv+mid\_adv)+UG\_TP(h\_unf+mid\_unf)] is depicted in blue. In order to give a comparison term, the red blobs indicate the activation for the UG\_MS (MPFC+AI/IFG). In the lower part of the figure, the graph (x=fairness level; y=beta-values; blue bars=MS; green bars=TP; yellow bars=FW; error bars=SEM) shows the activation in MPFC [14 52 40]: there is a difference in b-values between the

h\_dis and h\_adv in UG\_MS, but not in UG\_TP. In FW, despite the same trend, this activation did not reach significance.

## **5.4 Discussion**

In this study, I employed a modified version of the UG task, which allowed manipulating offers both for fairness and for advantageousness levels. The MS-TP distinction shows that the activation of left posterior superior and middle temporal gyri (BA 22) is higher in MS as opposed to TP. It is known from the literature that the both STG and MTG are involved in many kind of tasks, especially as far as semantic processing is concerned (Friederici, Opitz, and von Cramon, 2000; Luo et al., 2003); however, these areas play an important role in other processes, such as detecting biological motion (Adolphs, 2003), explaining and predicting the behavior of others during theory of mind scenarios, and also deciding about complex ethical dilemmas (Fletcher et al., 1995; Heekeren et al., 2003; Paulus, Feinstein, Leland, and Simmons, 2005). In the case of this study, the recruitment of this area in MS could reflect the fact that participants think about the other person's (A) situation, comparing it to his or her own situation. In TP, no comparison is going on because, probably, participants decide on the offers without putting themselves in the shoes of the third parties. The fact that, nevertheless, they reject unequal splits, suggests that the rejection of inequality should not be thought as an empathic response, but more as a heuristic that might be used by our cognitive system to face this kind of situation, when we have no other cues to drive decisions.

As far as our initial hypotheses are concerned, the results show a different activation of anterior insula (AI) on one side and of the ACC, extending to a more anterior part of the medial prefrontal cortex (MPFC) on the other side. In particular, AI showed a major activation for

extremely (and mid) unfair offers as opposed to fair offers, irrespectively of the target of the offers, whereas ACC and MPFC are active when facing unfair offers for myself, but not on behalf of third party; in particular, the anterior part of MPFC is more active when disadvantageous divisions are considered. These results appear to support our initial hypothesis, which claim the involvement of AI and MPFC in two different cognitive processes.

#### *5.5.1 Self-Specific Neural Networks*

A review of the literature about the functions of the different parts of the medial prefrontal cortex can be found in the discussion section in chapter 3; this area is accounted for self-other judgments, as well as for moral judgments, and, more broadly, for heterogeneous functions, which involve the co-occurrence of cognitive, emotional and social processes. As already reported, overlaps between these different cognitive processes have often been suggested (Jenkins et al., 2008, Hynes et al., 2006; Shamay-Tsoory et al., 2006; Gilbert et al., 2006), also in cause of the fact that some of these processes are confounded one another such as self-referential and emotional processing (Amodio and Frith, 2006). In the study described in chapter 3, I have reported an activation of the middle-anterior portion ( $y \approx 58$ ,  $z \approx 8$ ) of the MPFC (Figure 3.4b, green cluster) which, in contrast to the more ventral part involved mainly in self-perception, may reflect a recruitment of this area when facing self-directed unfair behavior, correlating this activation with emotional arousal evoked by a self-affecting unfair offer. As previously pointed out, the model of MPFC activation proposed by Amodio and Frith (2006) suggests that value-related representations extend the more anterior, the more complex they become, integrating with socio-affective processes. In accordance with this idea, in the current study, different portions of the MPFC were involved in different processes. In particular, two clusters within the medial

prefrontal cortex are differentially activated. On the one hand, the ACC activation ( $y \approx 18$ ) reflected the degree of unfairness of the offers: the more the offer was unfair, the more the area was active, irrespectively of the advantageousness of the offer. This U-shape activation was significant in UG\_MS specifically; it showed the same trend also in UG\_TP, nevertheless without reaching significance. It is known from the literature that cognitive demand requires the involvement of ACC (Pardo, Janer, and Raichle, 1990; van Veen, Cohen, Botvinick, Stenger, and Carter, 2001; Botvinick, Cohen, and Carter 2004), and these data go in this direction: the ACC is more active when the unfairness is higher, namely when the decision is harder and requires more cognitive effort, elicited by the contrasting principles that vie for the response. The fact that the activation is significantly high in MS but not in TP, despite showing the same trend in the latter, is plausible considering that taking a decision that affect the personal payoff is more effortful as compared to taking a decision that affects others.

On the other hand, a more anterior part of the medial prefrontal gyrus ( $y \approx 52$ ,  $z \approx 40$ ) was involved specifically in the perception of unfair disadvantageous offers; again, no significant effect for the UG\_TP was showed. In line with predictions of Amodio and Frith's model, whereas the more posterior part of the MPFC was implicated in actions and conflict monitoring, the more anterior part is associated with meta-cognitive representations and integrates the values of the possible outcomes with socio-affective motives: in this case, the higher activation related to disadvantageous offers, as opposed to fair, is likely to reflect the negative affective reaction to unfairness when it is specifically affecting the self. This result is in line with our previous study, in which the anterior part of the MPFC ( $y < 4\text{mm}$ ) was recruited when reacting to self-directed unfairness.

### *5.5.2 Fairness-Specific Neural Networks*

As far as the fairness, or equality, as it should be called in this case, perception is concerned, the hypotheses have been supported by the data. The activation of the cluster involving anterior insula, with an extension to the inferior frontal gyrus, bilaterally, was higher when processing unequal offers, as opposed to equal ones, irrespectively of the target. This activation could be interpreted as an involvement of the AI in processing general unfairness, rather than being correlated to the negative emotions elicited by the self-affecting unfairness, as proposed by Sanfey et al. (2003). In chapter 3, the results described AI as involved in reaction to the perceived unfairness; however, first of all it was not possible to disentangle the reaction to unfairness perceived as a general violation of a social norm from reaction to unfairness perceived as a damage both for the self and for a third-party, who was likely to be considered as a group member. Secondly, it was not possible to understand whether the activation of the AI was to be related either to the perception of the inequality of the outcome or to the perception of the unfairness of proposer's behavior. In this study, both issues have been addressed, diminishing the degree of identification with the third party and, most importantly, ruling out the perception of unfairness in relation to the behavior, focusing only on the inequality of the outcome. The results suggest that AI's activation reflects the inequality of the outcome rather than the violation of a socially accepted behavior, therefore, it is possible that the AI signals a deviation from the expected outcome, which, in this case, is to be considered the equal division. This interpretation is in line with Güroğlu et al. (2008), who found a higher AI's activation when participants (i.e. responders in the UG) engaged themselves in behaviors which deviated from the expected ones, such as accepting unfair offers or rejecting unfair offers when the unfairness was unintended; the authors hypothesized that AI correlates with the detection of norm violations, which are context-

dependent. My interpretation, in lights of the findings I have presented so far, is that AI's involvement in detecting violations is not limited to behavioral violations (i.e. unfair treatment), but extends also to the outcomes; in this case, the outcome was decided by chance, and still AI was active for unequal divisions.

However, despite this evidence, it is not clear whether the AI signals the deviation from equality considered as the desired outcome (moral principle), or as the expected outcome (normative rule). Further work is needed in order to clarify this issue, such as disentangle the unequal division from the perception of the normative, and expected, outcome.



## Chapter 6

### General discussion

Over recent years, plenty of evidence has suggested a link between emotions and people's judgments about the wrongness of certain behaviors, such as those which cause harm to other people (Haidt, 2003; Shaun, 2004; Moll et al., 2008), thus leading to the idea that these two aspects might be causally related, i.e. emotions, specially negative emotions, drive people to refuse the wrongness of behaviors (Greene et al., 2004; Koenigs et al., 2007)<sup>5</sup>. This account has been extended from the moral thinking to the field of economics; given that experimental data challenged the model of *homo economicus*, in that generally people tend to behave against their self-interest when facing fairness norms' violations (e.g. for the case of UG, see Guth et al., 1982), it has been hypothesized that the cause of the rejection of these violations had to be found in the emotional reactions (Pillutla and Mournighan, 1996). Neuroscientific evidence supporting

---

<sup>5</sup> This account is an extension of the somatic-marker hypothesis, originally formulated by Damasio et al. (1991), which proposes that emotional mechanisms can guide our behavior when facing complex and conflicting choices that cause our cognitive (as opposed to emotional) system to be overloaded and not sufficient anymore.

this account has been brought forward by Sanfey et al. (2003), who identified in the activation of the anterior insula the proof of an emotional involvement in fairness judgments, by van't Wout et al. (2006), who correlated the increase in the skin conductance response to the rejections of unfair splits, and by Koenigs et al. (2007), who found that patients with a ventromedial prefrontal damage, known for entailing deficits in emotional control, rejects more unfair offers as opposed to controls, just to cite three among the several studies on the topic.

In its traditional formulation, the UG, however is a self-centered task, and does not allow to disentangling between the perception of pure unfairness from the perception of self-affecting unfairness; it follows that it is not clear whether emotions have to be related to one aspect or the other of the task. At the beginning of this chapter (6.1) I will address this issue in lights of the results I have obtained in my studies.

Emotions have been linked to the rejection of unfairness, but what do people exactly reject, when rejecting an unfair division? Many accounts have been proposed to answer to this question. Rabin (1993) argued that people want to punish unfair behavior, considered a violation of the fairness norm, and for this reason they reject, even though this means to give up a certain gain. The *negative reciprocity account* has been supported by much evidence that described intentionality as necessary for the perception of unfairness and, consequently, for rejections (e.g., Blount, 1995, Güroğlu et al., 2008). Fehr and Schmidt (1999) proposed that people are more concerned about the inequality of the payoffs and, therefore, they are interested in keeping the payoff as equal as possible, by rejecting unfair divisions. In its formulation, however, the theory predicts a preference for self-centered inequality, which means that people are actually concerned about unequal splits, but they are much more prone to accept them when advantageous for themselves. All these accounts have been put forward in order to explain responders' behavior in

the UG. However, the preferences expressed by people playing the traditional UG do not allow to clearly supporting one of these theories. In fact, first of all in the UG there is always an involved proposer, so it is not clear whether people would reject even if this would not mean to punish the source of unfairness. Secondly, in the traditional UG as it is, it is not possible, in cause of the low plausibility, to administering participants with unfair but advantageous offers, making impossible to verify Fehr and Schmidt's prediction of a preference for self-centered inequality. These issues will be addresses, in lights of the results obtained, in paragraph 6.2.

### *6.1 Emotional involvement and self-concerns*

It has been widely discussed how responders in the UG are frustrated by the unfairness, and, as a cause of this, they irrationally reject unfair offers: the most established view is that the perception of unfairness, intended as a norm violation, elicits a negative emotional reaction. I have challenged whether this assumption holds even when unfairness is not self-affecting. To do so, I have developed a modified version of the UG in which participants were required to accept or reject offers either for themselves (myself condition) or on behalf of an unknown third party (third party condition). In this paradigm, responders had to decide both on unfair divisions that were affecting their own payoff (self-affecting unfairness) and on unfair splits that had absolutely no consequences for their payoff (pure unfairness). Whereas no difference in the rejection rate was found between myself and third party, both psychophysiological and imaging results suggest that signs of emotional involvement were actually correlated with unfairness, but only when this was affecting the self.

How can neuroscientific findings help to understand this decisional process? The fact that behavioral data does not show any difference between the two targets, while psychophysiological

and imaging data do, indicates that there is not a direct correlation between emotions and rejections. Even if it is not possible to claim that no emotion is involved in the third party condition, it is reasonable to argue that different mechanisms are involved in the two tasks; one mechanism, that involve the activation of the anterior MPFC, is recruited when reacting to self-directed unfairness, a condition which also elicit high emotional arousal, while the other mechanism, involving the AI, is recruited when reacting to pure unfairness, that is irrespectively of the target of the unfairness. As far as the MPFC is concerned, its selective activation when participants react to self-directed unfairness is in line with Amodio and Frith's model (2006), which claims that value-related representations in the ventral MPFC extend the more anterior (and superior), the more complex they become, and that they integrate with socio-affective processes. This model applies both to the results of chapter 3 and of chapter 5, where is shown that a further anterior area of the MPFC is more sensitive to unfairness when is disadvantageous for the subject.

The results suggest that there is something peculiar about the self-involvement in the perception of unfairness; the interpretations proposed so far in the literature, however, have generalized these peculiarities to a broader interpretation of the perception of norms' violations, by claiming that the rejection of norms' violations implies a negative emotional reaction which lead to irrational behaviors, such as in the UG case. In contrast, I argue that the emotional involvement is a side effect, which depends upon self-involvement. Thus, negative emotions are not the best candidate to explain exhaustively the aversion towards fairness violations; what about the other candidates, which are negative reciprocity and inequality aversion?

## *6.2 Negative reciprocity, inequality aversion and self-involvement as salient contextual cue.*

As discussed above, and extensively in chapter 4, the “involved proposer” is a limitation of the traditional UG, if we are interested in testing both the negative reciprocity and the inequality aversion account. The former has been tested by a number of studies, which found that manipulating the intentions of the proposer influenced the rate of rejections of responders (Blount, 1995; Güroğlu et al., 2008). However, in most of these studies, participants’ attention was focused on the intentions, leading them to take this variable into consideration (see chapter 4). Moreover, in some studies (e.g., Blount, 1995, Sanfey et al., 2003, in van’t Wout et al., 2006), when participants are required to decide upon offers made by a computer, they reject less, as compared to the condition in which the opponent is a human being, but they do not reject zero, as it would be expected if rejections are delivered as a punishment for proposer’s unfair behavior. The effects produced by the manipulations of the UG described in chapter 4 suggest that the will to punish an unfair proposer is not necessarily the only motive which drives rejections. As for the emotional involvement, I am not claiming that negative reciprocity does not contribute to the observed behavior but that this does not explain the data I presented here, given that participants rejected unequal division also when the allocation was decided by an external uninvolved proposer, or even by a computer algorithm.

In light of these results, it is evident that people reject the inequality of the outcome, even excluding the intended unfairness of the proposer. The second manipulation that I have introduced, made possible by the non-involvement of the proposer, focuses on the perception of inequality in relation to the self. The hypothesis, following Fehr and Schmidt’s account, is that responders are much more willing to accept unfairness when it is self-advantageous, and, in fact, it is exactly what happened: responders rejected inequality that was disadvantageous but accepted advantageous unequal splits. Interestingly, when asked to decide on behalf of

completely unrelated third parties, and when they had no reason to favor one person over the other, participants rejected unequal splits. Imaging data, discussed in chapter 5, corroborate the idea of different mechanisms involved in equality perception, whether it is self- or other-directed, and converge with the data discussed in chapter 3. On the one hand, MPFC is recruited specifically when the self is involved, in particular when the self is exposed to unfairness; moreover, an anterior part of the MPFC is more sensitive to disadvantageous unfairness as opposed to advantageous, suggesting the higher, probably emotional, salience of this condition. The activation in AI, on the other hand, is modulated by the degree of inequality, no matter whether it is self- or other- directed, or advantageous or disadvantageous for the responder.

There are accounts which suggest that our choices, and, consequently, our preferences regarding, for instance, monetary utility, vary with our sensitivity to the contingent contextual cues (Bicchieri, 2006). For this reason, we might decide to accept an unfair offer when we know that the proposer has no other choice, and reject it otherwise; similarly, we might accept an unfair offer, or even an unequal split, when we know that our opponent earned the right to be, to a certain extent, the winner (Camerer & Thaler, 1995). Taken together, the results presented here suggest that the self-involvement has to be considered one of these –very basic- salient contextual cues; in fact, when ruling it out, such as I did in the third party condition, people show a clean, perfect, U-shaped inequality aversion effect, suggesting that the self-involvement is affecting one side of the curve, pushing rejections towards the zero when offers were advantageous. The AI activation suggests also that a basic mechanism exists to signal this norm's deviation, and that an additional recruitment of the MPFC is needed when the salient cue is added. In the next paragraphs I will discuss these issues more in details.

### 6.3 *Inequality aversion as a default mode: moral norm or social heuristic?*

A quite heated cross-discipline debate on the role played by moral and social norms on behavior is interesting many researchers in different fields, such as economics, psychology, philosophy; Cristina Bicchieri, in her book *Grammar of Society* (2006), developed an account in which she has described people's behavior as guided by their knowledge of norms, and these norms get triggered depending on which cues are salient in a context. So, for instance, if the proposer has earned the right to be the proposer, this might not trigger the same norm that gets triggered in the standard UG, in which behavior is guided by non-better specified norms of fairness. More generally, how people behave in economic games, as in life, depends on which norms get triggered, and this depends on how people interpret the situation.

Here I suggest that the phenomenon of the inequality aversion has to be considered as a default mode; if there are no particular motives to prefer one person over the other, the unfair or unequal division is mostly rejected, even when rejections are not intended to signal a violation. This default mode is perturbed by salient environmental cues that shift the preferences towards one extreme or the other. These cues can be self-involvement, as we have seen so far, merit, or information that we have about our opponent (Camerer & Thaler, 1995); if we know that the proposer is a person in need, we might be a bit more tolerant towards unfair divisions, accepting more. As far as the original interpretation of inequality aversion, by Fehr and Schmidt, is concerned, I would propose, instead, to change the name to the definition; if a person rejects inequality only when it is self-disadvantageous, but accept it when self-advantageous, I would talk about competitiveness rather than inequality aversion.

Coming back to Bicchieri's account, another crucial distinction is the one between personal and social norms: according to the author, if the desire to conform to a norm is

conditional to our expectations on others following the same norm, then that norm is social. On the contrary, personal norms, which include moral norms and habits, such as, in Bicchieri's example, brushing one's teeth at night, do not depend on my expectations on others' behavior. The hypothesis is that equal-division norm, and, in general, fairness norms which are followed in the UG, are social norms; If we put people in the UG and we manipulate their expectations about whether others would divide equally, this should affect the offers that proposers make. In a recent set of studies, this is exactly what happened. Bicchieri and Xiao (2008) had participants play the Dictator game; before the participants played the game, they were set up with different expectations. One group was told: '60% of the dividers who participated in a session of this experiment last year approximately maximized their own earnings (i.e., their counterpart got 20% or less).' Another group was told: '60% of the dividers who participated in a session of this experiment last year shared the amount approximately equally (i.e., their counterpart got 40% or more).' What the authors found was that among participants who expected others to divide unequally only 33% conformed to the equal-division norm, whereas 52% of those who expected others to divide equally conformed to the equal-division norm. However, as discussed by Nichols (forthcoming), this evidence does not necessarily refute the idea that equal-division is a personal norm, if, free from expectations' biases, we desire to divide equally; nevertheless, as the authors claim, the desire to divide equally is not my only desire, given that, for instance, I also want to make money. As I have discussed in chapter 4, there are three principles that might drive behavior in UG: maximization of the payoff (rationality), competitiveness (self-oriented fairness) and pure fairness preference. Environmental cues might drive the choice: when the responder is personally involved in the bargain, the three principles vie to the response; when she is not personally involved, rationality principle and fairness are competing for the response. In light of



the previous considerations, not only environmental cues, but also personality traits are likely to be involved in the decisions; it might be the case that equal division is a personal norm for many people, but for many other people, equal division is, instead, a social norm. It is known from the literature that individual differences in performance in economic games can be explained by identifying people as prosocial, when they are more likely to seek equality in outcomes in economic games, individualists, when they care only about their self-interest, and competitive, when the interest is the maximization of the difference in the outcomes (Van Lange, 1999; Haruno & Frith, 2010). For this, I suggest here that accounts which consider the integration of all these aspects, i.e. default mode choices, personality traits and expectations, are the best candidates to explain behavior in economic games (Rustichini, 2009).

#### *6.4 The role of the anterior insula*

As I have discussed extensively in chapter 3, the AI plays a role in many different tasks involving different cognitive and emotional demands. A very detailed review has been published by Craig in 2009. AI is involved in interoception, pain perception and empathy, self-recognition, vocalization and music, emotional awareness, time perception, perceptual decision making, cognitive control and performance monitoring, etc. The author proposed an account that integrates all these findings by interpreting AI as involved in awareness, defined as “knowing that one exists”. In many studies reviewed, the findings might be interpreted as AI’s activation reflects deviation from the expected outcome, even though the expectation is not explicit. Just to cite a few examples, as far as cognitive control is concerned, AI’s activation is higher when perceiving the stop-signal (a signal that should trigger the inhibition of the response which was, in turn, previously triggered by the target); moreover, the activation increased together with the

rarity of the stop-signal (Brass, & Haggard, 2007; Ramautar, Slagter, Kok, and Ridderinkhof, 2006). As far as music perception is involved, Platel et al. (1997), using PET, found differentiated brain regions associated with familiarity, pitch, rhythm and timbre in music listening and in particular a robust activation in the left AIC during the rhythm task, in which occasional notes in well-known melodies were mistimed. Under this perspective, the self-awareness hypothesis can be interpreted as the integration of external stimuli with the complex set of motivations and expectations of the subject.

Here, I proposed that AI might signal either a deviation from fairness or, more in general, a deviation from an expected outcome. Following the arguments on norms presented above, AI could either signal the deviation from a desired outcome, if fairness is actually considered as the desired outcome (personal norm), or a deviation from the expected outcome (social norm). In order to disentangle between these two options, further work is needed: it is necessary to test AI's activation when modifying people's expectation in the UG task I have used in the experiment described in chapter 5, by shifting it towards a more or less equal outcome. If the AI's activation varies together with the shift of expectation, then it is likely AI reflects the social aspect of the equal-division norm; otherwise it could reflect the personal aspect. However, in any case, personality traits should not be neglected, and must be considered when evaluating expectations.

### *6.5 Concluding remarks*

Are human beings actually endowed with the desire to be altruistic and the will to divide resources equally, or do the altruistic behaviors that are performed depend on their will to conform to social norms, because they fear punishment or sanctions? The debate on whether

fairness should be considered as a moral or a social norm is vivid and of much interest. The importance of finding an answer to this question lies in the possibility to manipulate human cognitive and emotional reactions to certain events in order to increase behaviors such as, for instance, charity giving. How can neuroscience enter this debate? Neuroscientific data may be used to integrate, or to clarify, behavioral evidence, to help disentangling between different hypothesis, as in the case of the role played by AI.

On the other hand, how can the investigation of these topics help to understand of brain structures and functioning? Integrating the evidence from different studies, from those that use very basic tasks, such as perceptual decision making, to those that use more complex tasks, such as economic games, contributes to unfold the complexity of the functioning of brain areas, like in the case of AI. There is a lot of skepticism around this possible integration, both from hardcore economists and hardcore neuroscientists, as it is natural when new approaches start to question old and consolidate theories. However, the human being as we know it derives from a complex integration of functions, and it is this special integration that makes him/her unique. It is not possible to fully understand one aspect (e.g., economic choices) without taking in consideration another fundamental one (e.g., brain organization). Thus, the take-home message I would like to leave is that, assuming that the data are carefully interpret, and fast and easy enthusiasms are avoided, the road to go down to is the one of integrations among disciplines; and the better the communication among different professionals (economists, neuroscientists, philosophers and so on), the more reliable the results will be.

## References

- Abbink, K., Sadrieh, A., & Zamir, S. (1999). The covered response Ultimatum Game. Discussion Paper #191. The Hebrew University, Center for Rationality and Interactive Decision Theory.
- Adolphs, R. (2003). Cognitive neuroscience of human social behaviour. *Nat Rev Neurosci*, 4(3), 165-178.
- Amodio, D. M., & Frith, Chris D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nat Rev Neurosci*, 7(4), 268-277.
- Andrade, A., Paradis, A. L., Rouquette, S., & Poline, J. B. (1999). Ambiguous results in functional neuroimaging data analysis due to covariate correlation. *NeuroImage*, 10(4), 483-486.
- Aramaki, Y., Honda, M., Okada, T., & Sadato, Norihiro. (2006). Neural correlates of the spontaneous phase transition during bimanual coordination. *Cerebral Cortex*, 16(9), 1338-1348.
- de Araujo, I. E. T., Kringelbach, M. L., Rolls, E. T., & McGlone, F. (2003). Human cortical responses to water in the mouth, and the effects of thirst. *Journal of Neurophysiology*, 90(3), 1865-1876.
- Armin Falk, Ernst Fehr, & Urs Fischbacher. (2003). Reasons for conflict: lessons from bargaining experiments. *Journal of Institutional and Theoretical Economics JITE*, 159(1), 171-187.
- Atlas, L. Y., Bolger, N., Lindquist, M. A., & Wager, T. D. (2010). Brain mediators of predictive cue effects on perceived pain. *The Journal of Neuroscience*, 30(39), 12964-12977.
- Ashburner J.T., & Friston K.J. (2004) Rigid body registration. In R. S. J. Frackowiak, J. T. Ashburner, W. D. Penny, & S. Zeki, eds. *Human Brain Function* Academic Press, p. 653-655.
- Aumann, R. J. (2008). Game Theory. In S. N. Durlauf & L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics* (2nd ed., pp. 529-558). Basingstoke: Nature Publishing Group.
- Beauregard, M., Chertkow, H., Bub, D., Murtha, S., Dixon, R., & Evans, A. (1997). The neural substrate for concrete, abstract, and emotional word lexica a positron emission tomography study. *J. Cognitive Neuroscience*, 9(4), 441-461.
- Bechara, A., Damasio, H., Tranel, D., & Damasio, A.R. (2005). The Iowa Gambling Task and the somatic marker hypothesis: some questions and answers. *Trends in Cognitive Sciences*, 9(4), 159-162.
- Bechara, Antoine, & Damasio, Antonio R. (2005). The somatic marker hypothesis: a neural theory of economic decision. *Games and Economic Behavior*, 52(2), 336-372.

Bechara, Antoine, Damasio, Hanna, Damasio, Antonio R., & Lee, G. P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *The Journal of Neuroscience*, 19(13), 5473 -5481.

Bechara, Antoine, Damasio, Hanna, Tranel, Daniel, & Damasio, Antonio R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, 275(5304), 1293 -1295.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Bicchieri, C., & Zhang, J. (2008). “An embarrassment of riches: modeling social preferences in Ultimatum games”, in U. Maki (ed) *Handbook of the Philosophy of Economics*, Elsevier 2010

Bicchieri, C., & Xiao, E. (2008). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2), 191-208.

Blount, S. (1995). When social outcomes aren't fair: the effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, 63(2), 131-144.

Bolton, G. E. (1991). A comparative model of bargaining: theory and evidence. *The American Economic Review*, 81(5), 1096-1136.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *The American Economic Review*, 90(1), 166-193.

Bolton, G. E., & Zwick, R. (1995). Anonymity versus punishment in ultimatum bargaining. *Games and Economic Behavior*, 10(1), 95-121.

Botvinick, M., Nystrom, L E, Fissell, K., Carter, C S, & Cohen, J D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature*, 402(6758), 179-181.

Botvinick, M. M., Cohen, Jonathan D., & Carter, Cameron S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*, 8(12), 539-546.

Brass, M., & Haggard, P. (2007). To do or not to do: the neural signature of self-control. *The Journal of Neuroscience*, 27(34), 9141-9145.

Brett M., Anton J.L., Valabregue R., and Poline J.B. (2002) Region of interest analysis using an SPM toolbox. In Neuroimage Sendai, Japan.

Boucsein, W. (1992). *Electrodermal activity*, Plenum, New York, NY.

Calder, A. J., Lawrence, A. D., & Young, A. W. (2001). Neuropsychology of fear and loathing. *Nat Rev Neurosci*, 2(5), 352-363.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

- Camerer, C., & Thaler, R. H. (1995). Anomalies: ultimatums, dictators and manners. *The Journal of Economic Perspectives*, 9(2), 209-219.
- Camille, N., Coricelli, G., Sallet, J., Pradat-Diehl, P., Duhamel, J.-R., & Sirigu, A. (2004). The Involvement of the orbitofrontal cortex in the experience of regret. *Science*, 304(5674), 1167 - 1170.
- Charness, G., (1996). Attribution and reciprocity in a simulated labor market: an experimental investigation. Mimeo
- Chang, L. J., & Sanfey, A. G. (2009). Unforgettable ultimatums? Expectation violations promote enhanced social memory following economic bargaining. *Frontiers in Behavioral Neuroscience*, 3, 36.
- Chikazoe, J., Jimura, K., Asari, T., Yamashita, K.-ichiro, Morimoto, H., Hirose, S., Miyashita, Y., et al. (2009). Functional dissociation in right inferior frontal cortex during performance of go/no-go task. *Cerebral Cortex*, 19(1), 146-152.
- Ciaramelli, E., Muccioli, M., Làdavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, 2(2), 84-92.
- Civai, C., Corradi-Dell'Acqua, C., Gamer, M., & Rumiati, Raffaella I. (2010). Are irrational reactions to unfairness truly emotionally-driven? Dissociated behavioural and emotional responses in the Ultimatum Game task. *Cognition*, 114(1), 89-95.
- Le Clec'H, G., Dehaene, S., Cohen, L., Mehler, J., Dupoux, E., Poline, J. B., Lehericy, S., et al. (2000). Distinct cortical areas for names of numbers and body parts independent of language and input modality. *NeuroImage*, 12(4), 381-391.
- Coricelli, G., & Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, 106(23), 9163 -9168.
- Coricelli, G., Critchley, H. D., Joffily, M., O'Doherty, J. P., Sirigu, A., & Dolan, Raymond J. (2005). Regret and its avoidance: a neuroimaging study of choice behavior. *Nat Neurosci*, 8(9), 1255-1262.
- Craig, A. D. (2003). Interoception: the sense of the physiological condition of the body. *Current Opinion in Neurobiology*, 13(4), 500-505.
- Craig, A. D. B. (2009). How do you feel--now? The anterior insula and human awareness. *Nature Reviews. Neuroscience*, 10(1), 59-70.
- Craig, A. D., Chen, K., Bandy, D., & Reiman, E. M. (2000). Thermosensory activation of insular cortex. *Nature Neuroscience*, 3(2), 184-190.

- Critchley, H. D., Wiens, S., Rotshtein, P., Ohman, A., & Dolan, Raymond J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, 7(2), 189-195.
- Crockett, M. J., Clark, L., Tabibnia, G., Lieberman, M. D., & Robbins, T. W. (2008). Serotonin modulates behavioral reactions to unfairness. *Science*, 320(5884), 1739.
- Damasio, A. (2005). *Descartes' Error: Emotion, Reason, and the Human Brain*. Penguin (Non-Classics).
- Dana, J., Weber, R. A., & Kuang, J. X. (2006). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1), 67-80.
- Decety, J., Echols, S., & Correll, J. (2010). The blame game: the effect of responsibility and social stigma on empathy for pain. *Journal of Cognitive Neuroscience*, 22(5), 985-997.
- Denton, D., Shade, R., Zamarippa, F., Egan, G., Blair-West, J., McKinley, M., Lancaster, J., & Fox, P. (1999). Neuroimaging of genesis and satiation of thirst and an interoceptor-driven theory of origins of primary consciousness. *Proceedings of the National Academy of Sciences*, 96(9), 5304-5309.
- Dolcos, F., LaBar, K. S., & Cabeza, R. (2005). Remembering one year later: role of the amygdala and the medial temporal lobe memory system in retrieving emotional memories. *Proceedings of the National Academy of Sciences*, 102(7), 2626-2631.
- Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2), 268-298.
- Durston, S., Mulder, M., Casey, B. J., Ziermans, T., & van Engeland, H. (2006). Activation in ventral prefrontal cortex is sensitive to genetic vulnerability for attention-deficit hyperactivity disorder. *Biological Psychiatry*, 60(10), 1062-1070.
- Engelmann, J. B., Capra, C. M., Noussair, C., & Berns, G. S. (2009). Expert financial advice neurobiologically "Offloads" financial decision-making under risk. *PloS One*, 4(3), e4957.
- Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2), 293-315.
- Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences*, 11(10), 419-427.
- Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2), 63-87.
- Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868), 137-140.

- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3), 817-868.
- Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., & Frith, C. D. (1995). Other minds in the brain: a functional imaging study of “theory of mind” in story comprehension. *Cognition*, 57(2), 109-128.
- Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3), 347-369.
- Friederici, A. D., Opitz, B., & von Cramon, D Y. (2000). Segregating semantic and syntactic aspects of processing in the human brain: an fMRI investigation of different word types. *Cerebral Cortex*, 10(7), 698-705.
- Gilbert, S. J., Spengler, S., Simons, J. S., Steele, J. D., Lawrie, S. M., Frith, Christopher D, & Burgess, P. W. (2006). Functional specialization within rostral prefrontal cortex (area 10): a meta-analysis. *Journal of Cognitive Neuroscience*, 18(6), 932-948.
- Gneezy, U., & Rustichini, A. (2000). Pay Enough or Don't Pay at All. *Quarterly Journal of Economics*, 115(3), 791-810.
- Goel, V., Grafman, J, Sadato, N, & Hallett, M. (1995). Modeling other minds. *Neuroreport*, 6(13), 1741-1746.
- Greene, J D, Sommerville, R. B., Nystrom, L E, Darley, J M, & Cohen, J D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science (New York, N.Y.)*, 293(5537), 2105-2108.
- Greene, Joshua D, Nystrom, Leigh E, Engell, A. D., Darley, John M, & Cohen, Jonathan D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2), 389-400.
- Gu, X., & Han, S. (2007). Attention and reality constraints on the neural processes of empathy for pain. *NeuroImage*, 36(1), 256-267.
- Gul, F. (2008). Behavioural Economics and Game Theory. In S. N. Durlauf & L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics* (2nd ed., pp. 433-438). Basingstoke: Nature
- Güroğlu, B., van den Bos, W., Rombouts, S. A. R. B., & Crone, E. A. (2010). Unfair? It depends: neural correlates of fairness in social context. *Social Cognitive and Affective Neuroscience*, 5(4), 414 -423.
- Gusnard, D. A., Akbudak, E., Shulman, G. L., & Raichle, M. E. (2001). Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(7), 4259-4264.



Güth, W., Huck, S., & Müller, W. (2001). The Relevance of Equal Splits in Ultimatum Games. *Games and Economic Behavior*, 37(1), 161-169.

Guth, W. (1988). On the behavioral approach to distributive justice: A theoretical and experimental investigation, in S. Maital, editor, *Applied Behavioural Economics*, New York University Press, Vol. 2, p703-717

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4), 367-388.

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences*. Oxford: Oxford University Press.(pp. 852-870).

Harlé, K. M., & Sanfey, A. G. (2007). Incidental sadness biases social economic decisions in the Ultimatum Game. *Emotion*, 7(4), 876-881.

Haruno, M., & Frith, Christopher D. (2010). Activity in the amygdala elicited by unfair divisions predicts social value orientation. *Nature Neuroscience*, 13(2), 160-161.

Heekeren, H. R., Wartenburger, I., Schmidt, H., Schwintowski, H.-P., & Villringer, A. (2003). An fMRI study of simple ethical decision-making. *Neuroreport*, 14(9), 1215-1219.

Hertwig, R., & Herzog, S. M. (2009). Fast and frugal heuristics: tools of social rationality. *Social Cognition*, 27(5), 661-698.

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, Property Rights, and Anonymity in Bargaining Games. *Games and Economic Behavior*, 7(3), 346-380.

Hui, K. K., Liu, J., Makris, N., Gollub, R. L., Chen, A. J., Moore, C. I., Kennedy, D. N., et al. (2000). Acupuncture modulates the limbic system and subcortical gray structures of the human brain: evidence from fMRI studies in normal subjects. *Human Brain Mapping*, 9(1), 13-25.

Hynes, C. A., Baird, A. A., & Grafton, S. T. (2006). Differential role of the orbital frontal lobe in emotional versus cognitive perspective-taking. *Neuropsychologia*, 44(3), 374-383.

Jabbi, M., Swart, M., & Keysers, C. (2007). Empathy for positive and negative emotions in the gustatory cortex. *NeuroImage*, 34(4), 1744-1753.

Jenkins, A. C., Macrae, C Neil, & Mitchell, J. P. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proceedings of the National Academy of Sciences*, 105(11), 4507-4512.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). Opinion: the neural basis of human moral cognition. *Nature Reviews. Neuroscience*, 6(10), 799-809.

Kahneman, D., Knetsch, J. L., & Thaler, R. (1986). Fairness as a constraint on profit seeking: entitlements in the market. *The American Economic Review*, 76(4), 728-741.

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience*, 14(5), 785-794.

Kelly, A. M. C., Hester, R., Murphy, K., Javitt, D. C., Foxe, J. J., & Garavan, H. (2004). Prefrontal-subcortical dissociations underlying inhibitory control revealed by event-related fMRI. *The European Journal of Neuroscience*, 19(11), 3105-3112.

Kiebel S., & Holmes A.P. (2004). General linear model. In R. S. J. Frackowiak, J. T. Ashburner, W. D. Penny, & S. Zeki, eds. *Human Brain Function* Academic Press, p. 725-760.

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science*, 321(5890), 806-810.

Knoch, D., Nitsche, M. A., Fischbacher, U., Eisenegger, C., Pascual-Leone, A., & Fehr, E. (2008). Studying the neurobiology of social interaction with transcranial direct current stimulation—The example of punishing unfairness. *Cerebral Cortex*, 18(9), 1987 -1990.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, 314(5800), 829 -832.

Knutson, B., Bhanji, J. P., Cooney, R. E., Atlas, L. Y., & Gotlib, I. H. (2008). Neural responses to monetary incentives in major depression. *Biological Psychiatry*, 63(7), 686-692.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005a). Distributed neural representation of expected value. *The Journal of Neuroscience*, 25(19), 4806 -4812.

Koenigs, M., & Tranel, Daniel. (2007a). Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game. *The Journal of Neuroscience*, 27(4), 951 -956.

Koenigs, M., Young, L., Adolphs, R., Tranel, Daniel, Cushman, F., Hauser, M., & Damasio, A. (2007b). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446(7138), 908-911.

Kohlberg, L. (1969) Stage and Sequence: The Cognitive-Developmental Approach to Socialization. In *Handbook of Socialization: Theory in Research*, Ed. D.A. Goslin. Boston: Houghton-Mifflin.

Krajchich, I., Adolphs, R., Tranel, Daniel, Denburg, N. L., & Camerer, C. F. (2009). Economic games quantify diminished sense of guilt in patients with damage to the prefrontal cortex. *The Journal of Neuroscience*, 29(7), 2188-2192.

- Lamm, C., & Singer, T. (2010). The role of anterior insular cortex in social emotions. *Brain Structure & Function*, 214(5-6), 579-591.
- Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, 54(3), 2492-2502.
- Lamm, C., Nusbaum, H. C., Meltzoff, A. N., & Decety, J. (2007). What are you feeling? Using functional magnetic resonance imaging to assess the modulation of sensory and affective responses during empathy for pain. *PloS One*, 2(12), e1292.
- Lane, R. D., Fink, G R, Chau, P. M., & Dolan, R J. (1997). Neural activation during selective attention to subjective emotional responses. *Neuroreport*, 8(18), 3969-3972.
- Van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, 77(2), 337-349.
- Larquet, M., Coricelli, G., Opolczynski, G., & Thibaut, F. (2010). Impaired decision making in schizophrenia and orbitofrontal cortex lesion patients. *Schizophrenia Research*, 116(2-3), 266-273.
- Lee, G. P., Meador, K. J., Loring, D. W., Allison, J. D., Brown, W. S., Paul, L. K., Pillai, J. J., et al. (2004). Neural substrates of emotion as revealed by functional magnetic resonance imaging. *Cognitive and Behavioral Neurology*, 17(1), 9-17.
- Liebrand, W. B. G., Jansen, R. W. T. L., Rijken, V. M., & Suhre, C. J. M. (1986). Might over morality: Social values and the perception of other players in experimental games. *Journal of Experimental Social Psychology*, 22(3), 203-215.
- Luo, Q., Perry, C., Peng, D., Jin, Z., Xu, D., Ding, G., & Xu, S. (2003). The neural substrate of analogical reasoning: an fMRI study. *Cognitive Brain Research*, 17(3), 527-534.
- MacLeod, & MacDonald. (2000). Interdimensional interference in the Stroop effect: uncovering the cognitive and neural anatomy of attention. *Trends in Cognitive Sciences*, 4(10), 383-391.
- Malhi, G. S., Lagopoulos, J., Owen, A. M., Ivanovski, B., Shnier, R., & Sachdev, P. (2007). Reduced activation to implicit affect induction in euthymic bipolar patients: an fMRI study. *Journal of Affective Disorders*, 97(1-3), 109-122.
- McCulloch C.E., Searle S.R., Neuhaus J.M. (2008). *Generalized, Linear, and Mixed Models*. 2nd ed. Wiley-Interscience.
- Miller, E. K., & Cohen, J D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24, 167-202.

Mitzkewitz, M., & Nagel, R. (1993). Experimental results on ultimatum games with incomplete information. *International Journal of Game Theory*, 22(2), 171-198.

Modinos, G., Ormel, J., & Aleman, A. (2009). Activation of anterior insula during self-reflection. *PloS One*, 4(2), e4618.

Mohr, C., Leyendecker, S., & Helmchen, C. (2008). Dissociable neural activity to self- vs. externally administered thermal hyperalgesia: a parametric fMRI study. *The European Journal of Neuroscience*, 27(3), 739-749.

Moll, J., & de Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences*, 11(8), 319-321.

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., & Grafman, Jordan. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences*, 103(42), 15623 -15628.

Moll, J., De Oliveira-Souza, R., & Zahn, R. (2008). The neural basis of moral cognition. *Annals of the New York Academy of Sciences*, 1124(1), 161-180.

Moll, J., Zahn, R., de Oliveira-Souza, R., Bramati, I. E., Krueger, F., Tura, B., Cavanagh, A. L., et al. (2011). Impairment of prosocial sentiments is associated with frontopolar and septal damage in frontotemporal dementia. *NeuroImage*, 54(2), 1735-1742.

Moretti, L., Dragone, D., & di Pellegrino, G. (2009). Reward and social valuation deficits following ventromedial prefrontal damage. *Journal of Cognitive Neuroscience*, 21(1), 128-140.

Nichols, S. (2004). *Sentimental Rules: On the Natural Foundations of Moral Judgment*, Oxford University Press.

Nichols, S. (forthcoming). Emotions, norms, and the genealogy of fairness. *Politics, Philosophy & Economic*, 9(1) 1–22.

Nieuwenhuis, S., Holroyd, C. B., Mol, N., & Coles, M. G. H. (2004a). Reinforcement-related brain potentials from medial frontal cortex: origins and functional significance. *Neuroscience & Biobehavioral Reviews*, 28(4), 441-448.

Ochsner, K. N., Knierim, K., Ludlow, D. H., Hanelin, J., Ramachandran, T., Glover, G., & Mackey, S. C. (2004). Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience*, 16(10), 1746-1772.

Papeo, L., Corradi-Dell'acqua, C., & Rumiati, Raffaella Ida. (2011). “She” is not like “I”: the tie between language and action is in our imagination. *Journal of Cognitive Neuroscience*.

- Papeo, L., Vallesi, A., Isaja, A., & Rumiati, Raffaella Ida. (2009). Effects of TMS on different stages of motor and non-motor verb processing in the primary motor cortex. *PloS One*, 4(2), e4508.
- Pardo, J. V., Pardo, P. J., Janer, K. W., & Raichle, M. E. (1990). The anterior cingulate cortex mediates processing selection in the Stroop attentional conflict paradigm. *Proceedings of the National Academy of Sciences*, 87(1), 256 -259.
- Paulus, M. P., Feinstein, J. S., Leland, D., & Simmons, A. N. (2005). Superior temporal gyrus and insula provide response and outcome-dependent information during assessment and action selection in a decision-making situation. *NeuroImage*, 25(2), 607-615.
- Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *The Journal of Neuroscience*, 30(30), 10127-10134.
- Penny W.D., & Holmes A.P. (2004). Random effects analysis. In R. S. J. Frackowiak, J. T. Ashburner, W. D. Penny, & S. Zeki, eds. *Human Brain Function* Academic Press, p. 843-850.
- Peyron, R., García-Larrea, L., Grégoire, M. C., Costes, N., Convers, P., Lavenne, F., Mauguière, F., et al. (1999). Haemodynamic brain responses to acute pain in humans: sensory and attentional networks. *Brain: A Journal of Neurology*, 122 (9), 1765-1780.
- Pillutla, M., & Murnighan, J. K. (1996). Unfairness, Anger, and Spite: Emotional Rejections of Ultimatum Offers. *Organizational Behavior and Human Decision Processes*, 68(3), 208-224.
- Platel, H., Price, C., Baron, J. C., Wise, R., Lambert, J., Frackowiak, R. S., Lechevalier, B., et al. (1997). The structural components of music perception. A functional anatomical study. *Brain: A Journal of Neurology*, 120 (2), 229-243.
- Polezzi, D., Daum, I., Rubaltelli, E., Lotto, L., Civai, C., Sartori, G., & Rumiati, R. (2008). Mentalizing in economic decision-making. *Behavioural Brain Research*, 190(2), 218-223.
- Porro, C. A., Baraldi, P., Pagnoni, G., Serafini, M., Facchin, P., Maieron, M., & Nichelli, P. (2002). Does anticipation of pain affect cortical nociceptive systems? *The Journal of Neuroscience*, 22(8), 3206-3214.
- Preuschoff, K., Bossaerts, P., & Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron*, 51(3), 381-390.
- de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, 305(5688), 1254 -1258.
- Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 83(5), 1281-1302.

- Rainville, P., Duncan, G. H., Price, D. D., Carrier, B., & Bushnell, M. C. (1997). Pain affect encoded in human anterior cingulate but not somatosensory cortex. *Science*, 277(5328), 968-971.
- Ramautar, J. R., Slagter, H. A., Kok, A., & Ridderinkhof, K. R. (2006). Probability effects in the stop-signal paradigm: the insula and the significance of failed inhibition. *Brain Research*, 1105(1), 143-154.
- Rilling, J. K., Goldsmith, D. R., Glenn, A. L., Jairam, M. R., Elfenbein, H. A., Dagenais, J. E., Murdock, C. D., et al. (2008). The neural correlates of the affective response to unreciprocated cooperation. *Neuropsychologia*, 46(5), 1256-1266.
- Rustichini, A. (2009). Neuroeconomics: what have we found, and what should we search for. *Current Opinion in Neurobiology*, 19(6), 672-677. doi:16/j.conb.2009.09.012
- Saxe, R., & Powell, L. J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological Science*, 17(8), 692-699.
- Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, 300(5626), 1755 -1758.
- Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: an fMRI investigation. *Journal of Cognitive Neuroscience*, 18(5), 803-817.
- Schendan, H. E., & Kutas, M. (2011). Time course of processes and representations supporting visual object identification and memory. *Journal of Cognitive Neuroscience*, 15(1), 111-135.
- Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4(1), 25-55.
- Shamay-Tsoory, S. G., Tibi-Elhanany, Y., & Aharon-Peretz, J. (2006). The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. *Social Neuroscience*, 1(3-4), 149-166.
- Shapira, N. A., Liu, Y., He, A. G., Bradley, M. M., Lessig, M. C., James, G. A., Stein, D. J., et al. (2003). Brain activation by disgust-inducing pictures in obsessive-compulsive disorder. *Biological Psychiatry*, 54(7), 751-756.
- Singer, T., Critchley, H. D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences*, 13(8), 334-340.
- Steel, C., Haworth, E. J., Peters, E., Hemsley, D. R., Sharma, T., Gray, J. A., Pickering, A., et al. (2001). Neuroimaging correlates of negative priming. *Neuroreport*, 12(16), 3619-3624.

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., & Kirsch, P. (2011). Beyond revenge: neural and genetic bases of altruistic punishment. *NeuroImage*, 54(1), 671-680.

Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008). The sunny side of fairness. *Psychological Science*, 19(4), 339 -347.

Tomasino, B., Fink, Gereon R., Sparing, R., Dafotakis, M., & Weiss, P. H. (2008). Action verbs and the primary motor cortex: A comparative TMS study of silent reading, frequency judgments, and motor imagery. *Neuropsychologia*, 46(7), 1915-1926.

Tomasino, B., Werner, C. J., Weiss, P. H., & Fink, Gereon R. (2007). Stimulus properties matter more than perspective: An fMRI study of mental imagery and silent reading of action phrases. *NeuroImage*, 36(Supplement 2), T128-T141.

Tomb, I., Hauser, M., Deldin, P., & Caramazza, A. (2002). Do somatic markers mediate decisions on the gambling task? *Nature Neuroscience*, 5(11), 1103-1104.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., et al. (2002). Automated Anatomical Labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15(1), 273-289.

van Veen, V., Cohen, Jonathan D., Botvinick, M. M., Stenger, V. A., & Carter, Cameron S. (2001). Anterior cingulate cortex, conflict monitoring, and levels of processing. *NeuroImage*, 14(6), 1302-1308.

van't Wout, M., Kahn, R. S., Sanfey, A. G., & Aleman, A. (2006). Affective state and decision-making in the Ultimatum Game. *Experimental Brain Research*, 169(4), 564-568.

van't Wout, M., Chang, L. J., & Sanfey, A. G. (2010). The influence of emotion regulation on social interactive decision-making. *Emotion (Washington, D.C.)*, 10(6), 815-821

Weg, E., & Zwick, R. (1994). Toward the settlement of the fairness issues in ultimatum games : A bargaining approach. *Journal of Economic Behavior & Organization*, 24(1), 19-34.

Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. *Neuron*, 40(3), 655-664.

Winter, E., & Zamir, S. (2005). An experiment with Ultimatum bargaining in a changing environment. *Japanese Economic Review*, 56(3), 363-385.

Zamir, S. (2001). Rationality and Emotions in Ultimatum Bargaining. *Annals of Economics and Statistics / Annales d'Économie et de Statistique*, (61), 1-31.

Zysset, S., Huber, O., Ferstl, E., & von Cramon, D Yves. (2002). The anterior frontomedian cortex and evaluative judgment: an fMRI study. *NeuroImage*, 15(4), 983-991.